Méthodes de fouille de données en épidémiologie psychiatrique : application à l'analyse des facteurs et marqueurs de risque de la symptomatologie dépressive à l'adolescence.

Aminata Ali

Thèse de doctorat

NNT : 2021UPASR003

université PARIS-SACLAY

Méthodes de fouille de données en épidémiologie psychiatrique : application à l'analyse des facteurs et marqueurs de risque de la symptomatologie dépressive à l'adolescence

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 570, Santé Publique (EDSP)
Spécialité de doctorat: Santé publique-Epidémiologie
Unité de recherche : Université Paris-Saclay, UVSQ, Inserm, CESP, 94807, Villejuif, France
Référent : Faculté de médecine

Thèse présentée et soutenue à Paris, le 03 Mars 2021 par

Mme Aminata ALI

Composition du Jury

Emmanuel Chazard PU-PH, Université de Lille	Président
Maria MELCHIOR DR, INSERM	Rapporteur
Cédric GALERA PU-PH, Université de Bordeaux, ISPED	Rapporteur
Mohamed SEDKI MCU, Université Paris-Saclay (CESP, INSERM)	Examinateur
Bruno FALISSARD PU-PH, Université Paris-Saclay (CESP, INSERM)	Directeur de thèse
Barry Caroline IR (CESP,INSERM)	Co-encadrante de thèse
Catherine JOUSSELME PU-PH, Université Paris-Saclay (CESP, INSERM)	Invitée

Unité d'acceuil:

Ce travail doctoral a été réalisé au sein du Centre de recherche en Epidémiologie et Santé des Populations, INSERM U1018, Université Paris Saclay, situé au 97 boulevard de Port-Royal 75014 Paris, France.

Financement:

Ce travail de thèse a été financé par un contrat Cifre de l'ANRT, et a fait l'objet d'un contrat de collaboration entre l'unité INSERM U1018 et le pole universitaire de recherche du Centre Hospitalier Spécialisé Fondation Vallée.

Remerciements

Il *est naturel de commencer par ceux qui ont inspiré et dirigé ce travail.*

Je tiens à remercier chaleureusement Bruno Falissard. Merci d'avoir accepté d'être mon directeur de thèse et de m'avoir proposé de travailler sur ce sujet.

Je tiens à remercier ma co-directrice de thèse Caroline Barry, j'ai toujours mesuré la chance que j'ai eu d'effectuer ma thèse sous sa direction. Outre la profondeur de sa pensée scientifique, j'ai découvert quelqu'un de sincère, passionné et fort. Je te suis très reconnaissante de m'avoir accueillie et encadrée dès mon stage de Master 1. Je te remercie de la confiance que tu as pu m'accorder, de ta disponibilité, de tes qualités humaines, ainsi que de tout ce que j'ai pu apprendre auprès de toi. Merci Caroline pour toutes ses années passées à tes côtés, pour toutes nos discussions passionnées, pour ta pédagogie, ton soutien sur le plan professionnel comme personnel, ton écoute. Merci pour chaque mot et chaque paragraphe corrigé de la thèse, merci pour cette grande leçon de courage, merci d'avoir toujours été un exemple, merci pour tout…

Je tiens à remercier également Catherine Jousselme. Merci d'avoir accepté que je travaille à tes cotés et d'avoir permis que l'on obtienne ce financement de thèse. Merci de m'avoir accueillie avec entrain à la Fondation Vallée.

Toute ma gratitude va également aux membres de mon jury qui ont accepté d'étudier et d'évaluer ce travail doctoral.

Un grand merci à Christine Hassler et Caroline Huas pour tout votre soutien durant cette aventure. Merci pour vos relectures, corrections et pour toutes nos discussions. C'est auprès de vous et de mes directeurs de thèse que j'ai appris et je vous en serai très reconnaissante. Merci d'avoir été présentes tout au long de cette aventure, d'avoir su me remotiver, m'écouter et me soutenir surtout durant les derniers mois de thèse.

Merci à Soumaya pour ton amitié, ton soutien et nos beaux moments passées à l'unité : nos petits déjeuners, nos déjeuners, nos appels… Merci d'avoir été présente, de m'avoir conseillée et d'essuyer toutes ses larmes dans les moments difficiles.

Je tiens à remercier tous les membres de l'équipe U1018 ex U1178 encore présents ou non (Alexandra, Nour, Stéphane Bahrami, Florence Labrell, Massi, Sylvie, Juliette, Loubna, Constance, Laelia, Hugo basée à la Maison des Adolescents. Merci pour ces six années, merci pour votre accueil, la bonne ambiance qui y règne, votre gentillesse.

Merci à Mireille et Xu pour ces bons moments et nos longues discussions à la Fondation Vallée.

A ma famille,

Très tendrement à Aboubacar. Merci de me supporter au quotidien, merci pour ton soutien sans faille, merci pour ton amour.

A Dayan, sache que rien n'est impossible dans la vie et que le travail paye.

A Djibril, je sais que d'où tu es tu as pris soin de nous.

A mes parents, mes frères et sœurs, mes neveux et nièces j'espère que vous serez fiers de moi.

Valorisation scientifique de la thèse

Communication affichée

- « Intelligence artificielle en recherche biomédicale : une étude bibliométrique systématique » au colloque « Intelligence Artificielle et santé » organisée par la Faculté de Médecine Paris Sud le 25 novembre 2020.

Table des matières

Liste des tableaux

Liste des figures

Liste des Annexes

Liste des abréviations

ADRS : *Adolescent Depression Rating Scale*

ANN : Artificial Neural Network

CARET: *Classification And REgression Training*

CART: *Classification & Regression Trees*

DMML: *Data Mining/Machine Learning*

DSM-IV: *Diagnostic and Statistical Manual of Mental Disorders – 4ème édition*

DSM-V*: Diagnostic and Statistical Manual of Mental Disorders – 5ème édition*

ESCAPAD : Enquête sur la Santé et les Consommations lors de l'Appel de Préparation À la Défense

ESPAD: *European School Survey Project on Alcohol and other Drugs*

E-T: Ecart-type

HBSC: *Health Behaviour in School-aged Children*

IMC: Indice de Masse Corporelle

ICC : Coefficient de corrélation inter-classe

IC95%: Intervalle de Confiance à 95%

LASSO: *Least Absolute Shrinkage and Selection Operator*

MCMC : Monte-Carlo par Chaînes de Markov

OMS : Organisation Mondiale de la Santé

OR : *Odds Ratio*

RF: *Random Forest* (Forêt aléatoire)

Se: Sensibilité

Sp: Spécificité

SGD: *Stochastic Gradient Boosting* (Descente de gradient stochastic)

VPP : Valeur Prédictive Positive

VPN : Valeur Prédictive Négative

Avant-Propos

Méthodes et dépression de l'adolescent : construction de mon projet de thèse

L'unité Inserm 1178, s'intéressant à la santé mentale en santé publique, a de par son histoire une grande expérience des études sur les adolescents. De nombreuses études pour mieux connaitre les adolescents en population générale ou en population adolescente plus spécifiques ont été développées par cette unité. En 1993, l'« Enquête Nationale », s'est intéressée à l'adolescent dans sa globalité (1). Depuis lors, aucune autre enquête sur l'adolescent dans sa globalité n'a été mise en place. Les actions de préventions proposées et basées sur les données issues des précédentes études ne sont plus adaptées à l'adolescent d'aujourd'hui. En effet, « Les adolescents d'aujourd'hui vivent dans un monde qui bouge sans cesse, s'accélère, métisse les cultures, les repères, les dimensions » (2).

Afin de développer des actions de préventions adaptées et ciblées aux problématiques actuelles, il était nécessaire de mieux connaitre les adolescents dans une approche globale et non ciblée sur des comportements. Le pôle universitaire de recherche de la Fondation Vallée, dirigée par le Pr Jousselme et l'unité Inserm 1178, ont mis en place une enquête multicentrique en milieu scolaire s'intéressant au « processus adolescent » dans sa globalité explorant ainsi des domaines variés (santé physique, consommation de substances psychoactives, loisirs, relations familiales et aux pairs etc…), très proche de l'« Enquête Nationale ».

De cette enquête riche en thématiques explorées et en taille d'échantillon, mon sujet de thèse en santé publique a vu le jour et a donné lieu aux travaux qui vous seront présentés dans la suite de ce manuscrit. Mon sujet de thèse consistait en l'application des méthodes de fouilles de données (i.e data mining/ Machine Learning) à l'épidémiologie psychiatrique de l'adolescent. Afin de mener à bien ses travaux, un financement cifre de l'ANRT a été obtenu. Ce dispositif de financement a consisté en un contrat de collaboration entre la Fondation Vallée (considérée comme l'entreprise) et l'unité Inserm U1178 (considéré comme laboratoire d'accueil, rattaché à l'école doctorale de santé publique de l'université Paris Saclay).

Ainsi, mon travail de thèse est caractérisé par une double valence, d'une part méthodologique et d'autre part clinique. La valence méthodologique de cette thèse est qu'elle s'intéresse à l'utilisation des méthodes de fouilles de données. Pour cette partie j'ai été encadré par le Pr Falissard et le Dr Barry. Mon premier objectif a consisté à cartographier précisément l'utilisation réelle de ces méthodes en épidémiologie et santé publique. Par la suite l'intérêt de l'utilisation de différentes méthodes de Data Mining/Machine Learning dans l'analyse de l'association entre des variables explicatives et une variable d'intérêt clinique sur un jeu de données de santé publique/santé mentale a été évaluée. Pour

le versant clinique, l'analyse a porté sur l'association entre la symptomatologie dépressive à l'adolescence comme variable d'intérêt clinique et ses variables explicatives. L'objectif était de définir des sous-groupes d'adolescents à risque de présenter une forte symptomatologie dépressive. Cet aspect de ma thèse a été dirigé par le Pr Jousselme, professeur de psychiatrie de l'enfant et de l'adolescent à l'Université Paris-Saclay, encadrante en entreprise. La Fondation Vallée accueille principalement des jeunes enfants et adolescents en souffrance psychique. En clinique, les soins proposés par le Pr Jousselme sont multidimensionnels. Ils sont basés sur une prise en charge intégrative incluant au maximum les avancées scientifiques actuelles (neurosciences, épigénétique, approches neurocognitives, etc.) tout en conservant une vision psychodynamique du travail avec les enfants et leurs familles, permettant de s'adapter à l'histoire de chacun, quel que soit le cadre proposé.

1. Introduction

1.1. Contexte

1.1.1. L'adolescence

L'adolescence est connue pour être une période charnière du développement humain, se situant entre l'enfance et l'âge adulte, soit entre 10 et 19 ans selon l'OMS (2). Il s'agit d'une période reconnue comme sensible et principalement marquée par la puberté. La puberté est marquée à la fois par des changements physiologiques et psychiques, souvent interdépendants. Elle est accompagnée par des modifications concernant les relations familiales (plus particulièrement avec les parents) et les relations aux pairs. Les phénomènes physiques (pilosité, poitrine, mue…) bouleversent les représentations et les relations interpersonnelles (modifications cognitives). L'adolescence est une période cruciale pour le développement et la pérennisation d'habitudes sociales, importantes pour le bien-être mental (3).

Un développement identitaire avec des moments d'incertitudes et d'hésitations émergent également durant cette période de maturation. Les adolescents peuvent donc développer des personnalités provisoires (pouvant devenir définitives). Ils sont en quête d'identification à un groupe et de différenciation par rapport aux générations précédentes. Cet état de changement est une période de crise nécessaire pouvant être source de mal être et lieu d'apparition de trouble psychiatrique : le mal être pouvant aller d'un simple trouble du comportement à une tentative de suicide, non distinguable du travail psychique normal de l'adolescent. Tout facteur défavorable au développement à cet âge peut hypothéquer à long terme les possibilités d'épanouissement de l'adolescent, aboutissant à l'altération de la santé mentale telle que la dépression, tentative de suicide ou autre. Entre 10 et 20% des adolescents souffrent de problèmes de santé mentale dans le monde (4).

La plupart des troubles psychiatriques débutent précocement dans la vie: 50% avant l'âge de 15 ans et 75% avant 18 ans ; mais ils sont tardivement détectés et seule une minorité des jeunes en situation de souffrance psychique bénéficie de soins (5). En l'absence de soins, ces troubles évoluent de manière chronique. Ils occasionnent une souffrance psychique, des comorbidités addictives et somatiques (en particulier cardiovasculaires), une désinsertion sociale, une surmortalité précoce (suicide et somatique) et des coûts individuels, familiaux et sociétaux considérables, y compris à l'âge adulte (6–8).

1.1.2. La dépression à l'adolescence

L'adolescence est reconnue comme une réelle période de vulnérabilité pour la dépression, tant sur le plan psychologique que sur le plan biologique (9) .

La dépression est le problème de santé mentale le plus fréquent à cet âge. Ainsi, la prévalence ponctuelle de la dépression est selon les études comprise entre 0,4 % et 8,3 % (8% en France chez les 12-18 ans) (10,11).

Force est de constater qu'à l'adolescence, la dépression apparaît comme un état émotionnel complexe, impliquant de l'irritabilité, un sentiment d'être dépassé par l'expérience dépressive, une perception négative de soi, des pensées sur la mort. Il s'avère également qu'elle implique des manifestations non émotionnelles : lenteur mentale, troubles du sommeil, et manifestations par des interactions sociales à l'école, au travail, dans les loisirs et dans les relations avec les autres. Pour résumer, aucun symptôme n'est réellement spécifique à la dépression dans cette tranche d'âge et elle s'exprime davantage par des comportements et des somatisations (10).

Elle s'exprime différemment de chez l'adulte et ceci d'autant plus que, le sujet est jeune (12). Elle passe donc souvent inaperçue (5).

Plusieurs raisons sont à l'origine de cette sous-estimation :

- L'affect dépressif est fréquemment rencontré chez les adolescents en dehors de toute pathologie psychiatrique (9). La première difficulté consiste ainsi à distinguer ce qui relève de la dépressivité inhérente au processus même de l'adolescence de ce qui peut être le signe d'une évolution vers la pathologie. En effet, le caractère mouvant des processus psychiques durant cette période de transformations incessantes, couplé à l'absence de marqueurs d'évolution et au silence des intéressés, ne permet pas d'accéder à l'évolution des difficultés.
- Les critères diagnostiques cardinaux de l'épisode dépressif l'humeur dépressive persistante et/ou une perte de l'intérêt ou du plaisir - sont communs à l'adolescence et à l'âge adulte. Cependant, il existe des spécificités cliniques de la dépression à l'adolescence : la tristesse peut être remplacée par une irritabilité marquée (13), le ralentissement psychomoteur connaît de régulières levées transitoires, par exemple le temps d'une activité alors que les symptômes d'excitation motrice et d'agitation sont plus souvent présents que chez l'adulte dépressif. Enfin, les comorbidités psychiatriques sont plus fréquentes qu'à l'âge adulte : les comorbidités anxieuses (angoisse de séparation, trouble panique, phobies, 40 à 70 % des EDM, jusqu'à 40 % des dysthymies) ; les troubles du comportement (troubles des conduites, trouble

oppositionnel avec provocation de 20 à 80 %), surtout chez les garçons, ainsi que les consommations de substances psychoactives (de 20 à 30 %) (9).

- Certains adolescents présentent une dépression masquée, se manifestant par d'autres symptômes et conduites, se situant à première vue en dehors du champ sémiologique de la dépression. Par exemple des plaintes douloureuses et/ou symptômes somatiques (céphalées, règles douloureuses, douleurs dorsales ou abdominales, évanouissements …), des conduites d'apparence délinquantielle (passages à l'acte violents, transgressions, fugues, agressivité), des problèmes scolaires, des conduites à risques qui se traduisent par des accidents fréquents ou des conduites addictives. Ces « équivalents » dépressifs, distincts de la « crise d'adolescence », sont peu identifiés par les familles (11). Au final, l'agitation associée aux troubles du comportement divers orientent plutôt vers une prise en charge de type éducative, au risque d'omettre les aspects psychologiques, et donc méconnaitre la dépression.

D'un point de vue de santé publique, l'identification de caractéristiques cliniques précoces spécifiques permettrait à la fois de prévenir ou de réduire l'impact des troubles dépressifs. La dépression est un des troubles psychiatriques les plus fréquents, générant le plus de situations de handicap. Chez les adolescents, la dépression est d'ailleurs un facteur de risque majeur du suicide, seconde cause de décès dans ce groupe d'âge (14). Plus de la moitié des victimes de suicide dans cette période de la vie souffrait d'un trouble dépressif au moment de leur décès (15).

Comme le souligne le rapport Moro-Brisson, les soins en santé mentale des jeunes sont une priorité (16). Il est également important de noter que les différents éléments exposés en font une des priorités de santé publique du plan gouvernemental « ma santé 2022 » (17). Après avoir été longtemps banalisée à ce moment de la vie, la dépression de l'adolescent est devenue un vrai thème de préoccupation. Compte tenu des spécificités cliniques de la dépression à l'adolescence, des instruments de mesures adaptés ont été développés. Actuellement, une seule échelle, l'« Adolescent Depression Rating Scale » (ADRS), d'origine française, reconnue à l'international, a été validée pour quantifier l'intensité des symptômes dépressifs à l'adolescence et permettre une approche diagnostique (18).

Deux grandes études épidémiologiques françaises ont ainsi intégré l'ADRS dans leurs questionnaires. Ces deux enquêtes ont pour objectif de mesurer les consommations de substances psychoactives principalement : ESPAD 2007[1] « European School Survey Project on Alcohol and other Drugs » a inclus

[1] Géré par l'Observatoire Français des Drogues et Toxicomanie, répétée tous les 4 ans Site internet : https://www.ofdt.fr/

des adolescents scolarisés âgés entre 15 et 16 ans (19) ; ESCAPAD[2] « Enquête sur la Santé et les Consommations lors de l'Appel de Préparation À la Défense » a porté sur la santé des jeunes garçons et jeunes filles âgés de 17 ans(20).

Les prévalences de la dépression selon l'approche diagnostique de l'ADRS étaient plus élevées chez les filles que chez les garçons. Dans l'enquête ESPAD 2007, la prévalence de la dépression chez les filles était de 11,4% contre 5,4% chez les garçons (21) ; dans ESCAPAD 2008, la prévalence de dépression chez les filles était de 10,4% contre 4,5% chez les garçons (22) . L'ADRS n'a plus été utilisé depuis ni dans ESPAD ni dans ESCAPAD.

1.1.3. Variables explicatives et dépression à l'adolescence

Dans le paragraphe qui suit, je présente les caractéristiques associées (variables explicatives) à la dépression de l'adolescent dans la littérature. En 2015 (première année de thèse), Cairns et al ont publié une revue de la littérature systématique sur laquelle je me suis appuyée pour les répertorier. Ces caractéristiques associées à la dépression pouvaient être identifiées soit comme des facteurs[3]/marqueurs[4] de vulnérabilité soit comme des conséquences. J'ai complété ces informations, en particulier pour les variables explicatives dont la significativité n'a pas pu être montrée. Pour cela, j'ai effectué une recherche sur PubMed en utilisant les mots clés suivants : « Depression » AND « adolescent » et la caractéristique. La présentation suivante n'a pas vocation à être exhaustive mais à vous présenter les principales variables explicatives identifiées dans la littérature et/ou utilisées dans mon travail de thèse. Elles sont donc regroupées par thème, le sens de la relation est évoqué s'il est équivoque.

Parmi les variables explicatives répertoriées, on retrouve :

- Les données sociodémographiques et éducationnelles

Les données sociodémographiques et éducationnelles influent sur le risque de présenter une dépression à l'adolescence. La prévalence de la dépression varie selon le sexe et l'âge. Les filles présentent un risque de dépression plus élevé que les garçons (23). Les adolescents les plus âgés sont

[2] Géré par l'Observatoire Français des Drogues et Toxicomanie, répétée tous les 2 ans Site internet : https://www.ofdt.fr/

[3] Définition facteur de risque : paramètre, une pratique ou une caractéristique modifiable avec un lien supposé causal

[4] Définition marqueurs : paramètre non modifiable de l'environnement ou une caractéristique non modifiable d'un individu dont la présence s'accompagne d'une augmentation de la probabilité d'apparition d'un trouble sanitaire sans nécessairement de lien causal

plus à risque d'en présenter une que les plus jeunes. Les difficultés scolaires de type redoublement, déscolarisation sont également associées à la dépression (24).

- La situation familiale

Une situation familiale complexe a été identifiée comme associée à la dépression de l'adolescent, par exemple la séparation des parents, les placements itératifs, la négligence ou maltraitance, les conflits intrafamiliaux et le manque de soutien familial sont liés à un risque élevé de présenter une dépression (25). Le décès d'un des parents (plus particulièrement le suicide d'un des parents) augmente le risque de présenter une dépression (26). La dépression est plus fréquemment observée chez les adolescents dont les parents consomment des substances psychoactives ou ayant déjà eu une dépression ou un autre trouble psychiatrique (10).

- Le niveau socio culturel de la famille

Un faible niveau d'étude des parents , une faible catégorie socio-professionnelle des parents ou le chômage d'un des parents sont également associés à un risque élevé de présenter une dépression (27–29).

- Les comportements à risque

La consommation de substances psychoactives licites et illicites est associée à la dépression (30). La fréquence et la consommation d'alcool ont été identifiées comme facteurs prédictifs d'un niveau de dépression élevé. La consommation de tabac et cannabis sont associées à un niveau élevé de dépression chez l'adolescent. Ces consommations peuvent avoir des conséquences chez les adolescents (par exemple des difficultés scolaires ou conflit familial) qui, par la suite, participent à un risque accru de dépression. La consommation d'autres drogues (i.e ecstasy, MDA, champignons hallucinogènes) ainsi que la poly consommation ont également été identifiées comme variables explicatives de la dépression à l'adolescence. La littérature s'accorde à dire que la consommation de substances psychoactives peut à la fois être une cause ou une conséquence de la dépression (30).

Une sexualité à risque caractérisée par des rapports précoces dans la relation, non protégés, une interruption volontaire de grossesse ou une grossesse à l'adolescence, est également associée à risque de présenter une dépression (31,32).

La participation à des jeux dangereux de type « jeux du foulard » est également associée à un risque élevé de présenter une dépression (33) . Trois principaux types de jeux sont distingués (34) et repris par le ministère de l'Éducation Nationale (35) :

- Les jeux d'asphyxie, recouvrent toutes les pratiques d'auto ou hétéro asphyxie dans lesquelles les adolescents vont tenter de bloquer mécaniquement leur respiration le plus longtemps possible par des pratiques de strangulation (jeu du foulard), de suffocation (jeu du sac) ou d'apnée (jeu de la tomate). Ils ont été répertoriés internationalement et sont à différencier de l'automutilation, des intentions suicidaires. Selon les études, la prévalence varierait de 6% à 16% en France (36).
- Les jeux dits d'agression sont basés sur l'utilisation de la violence par un groupe d'adolescents sur un adolescent isolé. Ils peuvent être consentis ou subis, avec une inversion des rôles entre agresseurs et victimes. Ce ne sont pas des bagarres où il existe une motivation à la rixe (ex conflit de territoire).
- Les jeux de défis ou jeux de mort, s'appuient sur le principe du « cap ou pas cap » au travers duquel les adolescents se lancent des défis de plus en plus dangereux avec, parfois, le souhait que leurs exploits soient filmés et diffusés. Selon Romano et al., dans ces « jeux de mort » l'adolescent se met en danger par des conduites où la limite de la vie est sans cesse recherchée (traversée de route juste avant le passage d'une voiture, escalader de grandes hauteurs, etc.) (37) .

Le regroupement de ces catégories sous l'appellation « jeux dangereux » est essentiellement utilisée en France. Ce pourquoi dans la suite de ce travail l'appellation sera écrite entre guillemets.

- <u>Certaines Comorbidités</u>

La perturbation du rythme du sommeil est associée au risque de présenter une dépression. La littérature s'accorde à penser que la perturbation du rythme de sommeil est un symptôme de la dépression mais aussi que la dépression est une cause des troubles du rythme du sommeil.

Les troubles des conduites alimentaires, en particulier les restrictions alimentaires (tentative délibérée de réduction, de limiter la consommation afin de réduire le poids), sont également associés à un risque élevé de présenter une dépression à l'adolescence (30).

La présence d'une maladie chronique ou d'une maladie grave augmente le risque de dépression (10,38). De même pour les antécédents traumatiques tels que l'état de stress post traumatique, l'agression physique ainsi que les violences, qui augmentent la prévalence de la dépression de l'adolescent (10).

- Le rapport au corps et à l'alimentation

Un poids (ou Indice de Masse Corporelle IMC) élevé est également associé à un risque élevé de présenter une dépression. Par exemple, un poids élevé peut détériorer l'image de soi ou avoir des répercussions somatiques, entrainant alors une augmentation du risque de présenter une dépression à l'adolescence. De même, la dépression à l'adolescence, peut contribuer à une augmentation du poids (30,39).

- La Sexualité

Les études de Kann et al., et Marshall et al., ont mis en évidence qu'une attirance sexuelle non hétérosexuelle (i.e bisexuelle ou homosexuelle) est associée à un risque élevé de présenter une dépression (40,41). L'entrée précoce dans la sexualité évaluée est également associée à un risque élevé de présenter une dépression (42).

- La qualité des relations interpersonnelles

Une relation négative avec les pairs ou la famille apparait comme un facteur de risque de présenter une dépression, de même que des évènements de vie négatifs. Par exemple, la relation avec les pairs peut être évaluée par le nombre d'amis. L'étude de Ueno et al., en 2005, a montré que le nombre d'amis était associé à la dépression à l'adolescence : un nombre trop élevé ou trop faible d'amis augmente le risque de présenter une dépression (43).

Field et al., ont mis en évidence qu'une mauvaise relation avec les parents est associée au risque de présenter une dépression à l'adolescence (44). La relation avec les parents peut être évaluée dans les études par des questions sur le type de discussion que les adolescents ont avec les parents, sur la vie, les sentiments et les expériences.

- La pratique de loisirs et d'activités extra-scolaires

Plusieurs études se sont intéressées au lien entre la participation à des activités extra scolaires (par exemple la participation à des ateliers d'écriture ou le théâtre), et la dépression à l'adolescence. Elles ont montré que la participation à, au moins une activité extra-scolaire est associée significativement à un faible risque de présenter une dépression (45–47).

Plusieurs études se sont intéressées au lien entre la pratique d'une activité physique, et les symptômes dépressifs (48,49). Plus précisément, l'activité physique correspond à «tout mouvement produit par les muscles squelettiques, responsables d'une augmentation de la dépense énergétique» (50) . D'autres études se sont intéressées quant à elles au lien entre la pratique régulière d'un sport et le

risque de présenter une dépression à l'adolescence (51–53) . Par définition, le sport est une activité physique réalisée de manière organisée avec un cadre, des règles et parfois même des compétitions. La littérature s'accorde à dire que la pratique d'un sport ou d'une activité physique est associée à un plus faible risque de présenter une dépression durant l'adolescence (48,49,51–53) .

- L'intensité de l'exposition aux écrans

L'intensité d'exposition aux écrans (télévision, Internet, jeux vidéo) a été étudiée de nombreuses fois. Les résultats sont équivoques. Certains auteurs ont mis en évidence une association positive entre l'intensité d'exposition et le risque de présenter une dépression : Plus l'intensité est élevée plus le risque de présenter une dépression est élevé (54). D'autres, ont suggéré une association négative (55). Tandis que la méta analyse de Liu et al., a mis en évidence une relation significative non linéaire entre l'exposition aux écrans et la dépression à l'adolescence. Les adolescents jouant peu ou intensivement aux jeux vidéo sont plus à risque de présenter une dépression (56).

D'autres variables explicatives ont été identifiées dans la littérature: religion, harcèlement scolaire, violence… (30,57) . J'ai choisi de ne pas décrire ces variables car les données ne sont pas utilisées dans mon travail de thèse (non collectées dans l'enquête utilisée). Leur absence est bien évidemment discutée dans les points forts et limites de mon travail.

Malgré les nombreuses variables explicatives identifiées, la littérature est unanime pour conclure que la prédiction du risque de présenter une dépression à l'adolescence reste peu performante. En effet, les études n'ont jusqu'alors testé les variables explicatives qu'isolément.

Une des voies d'amélioration de la prédiction serait l'analyse systématique et approfondie des combinaisons/interactions entre toutes les variables explicatives. Très peu d'études ont pu analyser systématiquement les combinaisons/interactions entre les variables explicatives, notamment à cause de l'insuffisance de la taille des échantillons et/ou parce que les données ne recueillaient pas les différents types de variables ensemble.

La multiplicité des interactions potentielles à analyser soulève des problèmes méthodologiques. En santé publique, les techniques issues des méthodes de « fouille de données » : Data mining/ Machine learning semblent de plus en plus utilisées sur des problématiques similaires, par exemple, facteurs de risques cardiovasculaires (58), sans qu'on en connaisse la fréquence d'utilisation réelle. Les avantages distincts de ces approches sont particulièrement importants ici notamment par leurs capacités à gérer un très grand nombre de variables et de combinaisons/interactions potentielles à tester. Ces méthodes permettent également de modéliser des relations complexes entre prédicteurs, allant bien au-delà des modèles additifs, interactifs et linéaires traditionnels. Pour empêcher le sur-apprentissage, des

stratégies type validation croisée sont utilisées. Ainsi, ces techniques ont le potentiel de produire des modèles qui reflètent la nature complexe du risque de dépression (59). Les techniques de Data Mining/Machine Learning (DMML) sont décrites dans le chapitre 2 de mon travail.

Pour autant, la fouille de données a été peu appliquée en épidémiologie psychiatrique ; une réflexion méthodologique est donc nécessaire pour déterminer les méthodes les plus adaptées aux particularités de la psychiatrie.

1.2. Objectifs de la thèse

Nous nous sommes demandé si les méthodes de DMML présentaient un intérêt supérieur aux modèles de régression pour prédire la dépression à l'adolescence?

Mon travail de thèse s'est intéressé à l'application des méthodes issues de la fouille de données à l'épidémiologie psychiatrique et en particulier à la dépression durant l'adolescence. Il s'est articulé en trois parties : i) cartographier l'utilisation réelle de ces méthodes en épidémiologie et santé publique en utilisant une technique de DMML ii) évaluer l'intérêt des méthodes type DMML dans l'analyse de l'association entre des variables explicatives et la symptomatologie dépressive à l'adolescence, iii) Définir des sous-groupes d'adolescents à risque de présenter une forte symptomatologie dépressive.

Résumé du contexte et objectifs

La dépression à l'adolescence est fréquente, avec un risque d'hypothéquer à court, moyen ou long terme, la santé mentale, la santé physique et l'insertion socioprofessionnelle. Repérée et traitée précocement, ses conséquences pourraient être atténuées (6,60).

Compte tenu de la symptomatologie fluctuante et spécifique, elle est difficile à identifier et nombre de ces jeunes déprimés ne sont pas pris en charge. Les données sociodémographiques et éducationnelles, la situation familiale, la relation avec les pairs et la famille, les habitudes de consommation de substances psychoactives, le rapport au corps ou à l'alimentation, les loisirs/activités physiques ont été identifiés dans la littérature comme associés à la dépression à l'adolescence (30).

Afin de simplifier la lecture du manuscrit, le terme « variable explicative » regroupera l'ensemble des facteurs/marqueurs de risque et de protection de la dépression.

Objectifs de thèse :

i) Cartographier l'utilisation réelle de ces méthodes en épidémiologie et santé publique en utilisant une technique de DMML

ii) Evaluer l'intérêt des méthodes type DMML dans l'analyse de l'association entre des variables explicatives et la symptomatologie dépressive à l'adolescence,

iii) Définir des sous-groupes d'adolescents à risque de présenter une forte symptomatologie dépressive

2. Etat de l'art de l'utilisation des méthodes de DMML en épidémiologie et santé publique.

Durant les deux dernières décennies, de nombreux domaines biologiques et biomédicaux ont assisté à l'explosion de leurs ensembles de données, passant de quelques centaines de données collectées à des millions (61). Cette évolution a concerné particulièrement le domaine de la génomique et les autres domaines « omiques », mais pas seulement. De nos jours, les grandes bases de données épidémiologiques peuvent inclure des informations sociodémographiques, administratives, cliniques, moléculaires, comportementales, environnementales et même des données textuelles issues des réseaux sociaux. Les grandes études de cohorte contemporaines collectent des centaines, voire des milliers de covariables d'expositions environnementales et de mesures comportementales au niveau individuel et écologique (62). Bien évidemment, lorsque l'on considère les interactions, comme dans le cas des études gène-environnement, la dimensionnalité augmente encore plus rapidement. Cette dimensionnalité introduit de sérieux défis, en relation, entre autres, avec l'accumulation de bruit, les problèmes de tests multiples et les facteurs de confusion qui peuvent aboutir à de fausses inférences statistiques (63). Parallèlement, un débat est apparu dans le contexte de la « crise de la reproductibilité », sur l'utilisation de la "signification statistique" et de la p-valeur pour déterminer de manière binaire si une analyse est intéressante ou non (61). Jeff Leek, parmi d'autres statisticiens influents, a affirmé que « ces méthodes qui ont été développées pour un monde où les informations étaient rares et difficiles à collecter ne sont pas adaptées pour traiter des ensembles de données plus importants, plus diversifiées et plus complexes » (61).

2.1. Historique et définition du DMML

Le terme « Fouille de données » ou « Data mining » introduit au début des années 1990, a été adopté par les informaticiens pour décrire les méthodes visant à « découvrir des relations et des structures utiles dans des données qui n'étaient pas connues auparavant ». Ce terme étant à la mode, une grande variété de définitions et de critères plus ou moins consensuels ont depuis vu le jour (64). Au-delà des divergences de définitions, le data mining peut être caractérisé comme un processus inductif, par opposition à l'approche hypothético-déductive traditionnelle des statistiques (65). En effet, (i) la « fouille de données » est par essence guidée par les données (qu'elles soient de nature massive ou non) et (ii) de nombreux algorithmes utilisés ont leurs racines intellectuelles dans le Machine Learning (66). Par définition, ces algorithmes sont programmés pour apprendre à effectuer une tâche automatiquement à partir des données, plutôt que d'avoir un fonctionnement explicitement modélisé.

Il existe une multitude d'algorithmes de Data mining et Machine Learning (DMML). Les classer et les répertorier de manière exhaustive est un challenge en soi. Cependant, presque toutes les méthodes

peuvent être formulées comme de l'identification de structure dans les données (67). Le chapitre qui suit a pour objectif de résumer les différentes familles d'algorithmes.

En DMML, le vocabulaire est spécifique. Dans les algorithmes d'apprentissages supervisés, on parle de données labélisées pour désigner la variable d'intérêt clinique comme la symptomatologie dépressive. Sur la base d'un ensemble de prédicteurs (les variables explicatives), on entraîne les programmes à prédire ces labels (68). L'apprentissage supervisé comprend des centaines d'algorithmes pour des tâches de classification et de régression. L'entrainement est effectué à partir d'un échantillon « d'apprentissage » et la prédiction est évaluée sur un échantillon « test » et/ou par une validation croisée. Les algorithmes d'apprentissage non supervisés regroupent principalement des algorithmes de clusterisation, de réduction de dimensionnalité ou d'association. Le Tableau 1, tiré de l'article de Wiemken et al., 2019, présente les différences de vocabulaire entre les biostatistiques et le DMML (69).

Tableau 1: Correspondance des termes utilisés en épidémiologie et Machine Learning[5]

Term in Epidemiology and biostatistics	Term in machine/statistical learning
Dependent variable; outcome variable; response variable	Label/class
Independent variable; predictor variable; explanatory variable	Feature
Contingency table; 2*2 table	Confusion matrix
Sensitivity	Recall
Positive predictive value	Precision
Deep learning	Artificial neural network with more than 1 hidden layer
Outcome group with the highest frequency	Majority class
Outcome group with the lowest frequency	Minority class
Proportion of cases in each category of the outcome variable (when outcome is categorical)	Class balance

[5] Tableau tiré de l'article de Wiemken et al., « Table 1 : Linking terms phrases in epidemiology and machine learning »(69)

Parmi les algorithmes les plus populaires, on distingue :

- **Les algorithmes de régularisation** :

Ils sont une extension des algorithmes de régression. Ils permettent de s'affranchir des problèmes d'estimation des paramètres, et de sur-ajustement. Ils rendent plus facile l'interprétation des modèles en présence d'un grand nombre de covariables dans le modèle de régression. Une pénalité est imposée aux modèles sur le nombre de coefficients ; cela permet donc de conserver les covariables ayant une force d'association importante avec la variable d'intérêt. On compte parmi ces méthodes les régressions LASSO, Ridge et Elastic net (70).

- **Les machines à vecteurs de support :**

L'objectif principal est de classer des observations en fonction de leurs caractéristiques et leur distance par rapport aux autres. Les machines à vecteurs de support (ou support vector machine, SVM) projettent les données dans un nouvel espace et classifient les observations en trouvant l'hyperplan de séparation optimal entre les observations de labels différents dans l'espace des covariables (71).

- **Les arbres de décisions** :

Ils reposent sur un partitionnement récursif des individus et sont représentés par des arbres (ensemble de règles de décision construites par l'algorithme). On distingue deux types d'arbres : les arbres de régression qui permettent de prédire une variable continue ou catégorielle ordonnée et les arbres de classification qui permettent quant à eux de prédire une variable catégorielle non ordonnée. L'algorithme de classification le plus utilisé est l'algorithme CART (Classification And Regression Trees). Les arbres de décisions sont des algorithmes qui permettent de produire des classifications compréhensibles (72) .

- **Les algorithmes bayésiens :**

Il s'agit d'un ensemble de méthodes qui utilise le théorème de Bayes pour des problèmes de classification et de régression. Les algorithmes les plus célèbres sont les algorithmes de classification naïve bayésienne (73).

- **Les algorithmes de réseaux de neurones artificiels et d'apprentissage profond « deep learning »:**

Les réseaux de neurones artificiels et les algorithmes de « deep learning » sont des modèles informatiques inspirés du système nerveux central (74) . Les algorithmes de réseaux de neurones sont utilisés à la fois en classification (réseaux de Kohonen) et en prédiction (perceptron et réseaux à

fonction radiale de base). Les domaines d'applications sont nombreux : reconnaissance d'image, diagnostic médical ou encore traduction automatique.

- **Les algorithmes d'ensemble ou algorithmes d'agrégation :**

Ces algorithmes sont des méta-modèles composés de plusieurs algorithmes entrainés indépendamment et dont les prédictions sont combinées pour faire une prédiction globale (par exemple, les descentes de gradient stochastique (SGD), les forêts aléatoire (RF)). Ces méthodes reposent sur la construction d'un modèle prédictif combinant un grand nombre de modèles individuels : il s'agit donc d'une agrégation des prédictions obtenues indépendamment (72).

- **Les Algorithmes de réduction de la dimensionnalité :**

Ces algorithmes non supervisés exploitent les structures des données afin de résumer les informations ou de simplifier les données qui peuvent ensuite être utilisées dans un modèle de prédiction. L'analyse en composantes principales (ACP) permettant de réduire le nombre de variables et l'analyse discriminante linéaire (dont le but est de prédire l'appartenance d'un individu à un groupe à partir de covariables) en font partie (70).

- **Les algorithmes de clusterisation :**

Ces algorithmes regroupent un ensemble de méthodes qui utilisent les structures inhérentes aux données pour les organiser en groupes aussi homogènes que possible : les individus présentant des propriétés similaires sont regroupés au sein d'un même groupe lorsque la distance est faible. La méthode des K-means et la classification hiérarchique font parties de cette famille d'algorithme. La première permet de répartir les individus en k cluster définis *a priori*, à partir de la distance euclidienne entre les individus ; tandis que la classification hiérarchique, ne nécessite pas de définir le nombre de cluster attendu (70).

La plupart des algorithmes de clusterisation sont répertoriés dans les techniques d'apprentissage non supervisé c'est-à-dire que les clusters sont construits uniquement à partir de l'ensemble des covariables. Il existe tout de même quelques algorithmes supervisés de clusterisation tels que les régressions sur profil (Bayesian profile regression) qui ont pour objectif de créer des sous-groupes homogènes d'individus à risque de présenter une caractéristique d'intérêt (75).

Au final, il n'y a pas de frontière claire entre les modèles de DMML et les modèles statistiques traditionnels (76) . Le fait qu'une méthodologie donnée soit considérée comme relevant du Machine Learning ou « des statistiques» reflète parfois son histoire autant que de véritables différences. Les

méthodes de régularisation ou les arbres de classification et de régression peuvent être considérés ou non comme de l'apprentissage automatique selon la personne interrogée.

Les données textuelles constituent une classe de données particulière. Comme pour l'analyse de données numériques, l'analyse de données textuelles, permet d'extraire de la connaissance, en détectant des patterns, identifiant des relations entre les mots ou recherchant des similarités. La fouille de données textuelles (« text mining » dans la littérature anglophone), peut aussi être approximativement catégorisée en méthodes non supervisées et supervisées. Les premières, ont en général pour but de structurer le contenu d'un corpus de documents et d'en découvrir les thèmes abordés sans connaissance *a priori*. La bibliométrie en est un champ d'application. Les méthodes supervisées quant à elles, ont pour objectif de rechercher des règles permettant l'affectation automatique d'un document à un thème prédéfini parmi d'autres. De nombreuses étapes de traitement automatique du langage naturel sont nécessaires avant de pouvoir appliquer des méthodes probabilistes. Au final, il est donc possible d'extraire des connaissances inconnues, valides et exploitables dans des documents textuels via la mise en œuvre de méthodes de DMML.

2.2. Utilisation du DMML en épidémiologie/recherche biomédicale

Les approches DMML ont été appliquées à une grande variété de recherches épidémiologiques. À ce jour, la littérature abonde de revues systématiques synthétisant les applications des approches DMML dans des domaines spécifiques ; par exemple pour l'épidémiologie de la pollution de l'air, le diabète ou la santé mentale (77–80). À ma connaissance, il n'existe pas de revue systématique qui présente le tableau complet des applications du DMML en épidémiologie et/ou en santé publique. Les revues généralistes existantes sont des revues narratives qui décrivent la situation et précisent les orientations actuelles ou prometteuses en matière de santé publique à l'aide d'exemples tirés de la littérature (81). Presque toutes sont d'accord pour conclure que les domaines du Machine Learning se développent rapidement. Cependant, presque toutes concluent qu'il existe toujours des obstacles qui empêchent une plus grande intégration des méthodes de DMML en épidémiologie, tels que des barrières culturelles et linguistiques, des questions éthiques (82,83) ou parce que les modèles du Machine Learning sont souvent des « boîtes noires » dont l'opacité empêche l'interprétation (81). Les autres sujets de préoccupation concernent la qualité des grands ensembles de données utilisés et les risques de sur-ajustement. Au final, il reste donc difficile d'avoir une idée quantifiée de l'impact réel des méthodes DMML sur la recherche en santé publique en raison du manque de revues systématique quantitatives ayant pour objectif d'évaluer l'ensemble du domaine.

J’ai donc entrepris une étude bibliométrique, visant à fournir une description objective et quantitative de l'engagement de la recherche en santé publique dans les approches DMML. A travers un examen temporel, j’ai cherché à cartographier de manière dynamique le paysage du domaine. Plus précisément, les objectifs de cette étude étaient i) d'examiner la croissance globale de la production de ce type de recherche, ii) de caractériser les revues les plus engagées, iii) d'identifier les domaines de recherche en santé publique les plus investis par les approches DMML et iv) de décrire les trajectoires de développement de ces domaines au fil du temps. Pour aborder ces questions, j’ai choisi de modéliser la dynamique de publication dans la base de données PubMed au cours de la dernière décennie.

Le nombre d’articles en santé publique et/ou épidémiologie sur la dernière décennie était trop important pour une analyse manuelle. Ma thèse étant centrée sur l’application des méthodes de DMML, j’ai choisi d’utiliser une technique de modélisation thématique non supervisée qui est une approche par «text mining ». La modélisation thématique a pour but d’organiser comprendre, rechercher et résumer automatiquement des thèmes récurrents au sein d’un ensemble de documents et privilégie la sémantique à la syntaxe (présentée dans le paragraphe 2.3.2 page 31).

2.3. Matériel et méthode

2.3.1. Critères de sélection / extraction des données

J’ai effectué une analyse bibliométrique de la littérature scientifique de PubMed en santé publique et épidémiologie traitant des concepts de DMML. Les critères de recherche étaient les suivants : articles, non dupliqués, avec un résumé disponible en anglais, et publiés entre le 01 janvier 2010 et le 31 décembre 2019. Les commentaires, lettres, rapports de consensus, les notes cliniques, les articles sans résumés ou portant sur les animaux ou végétaux ont été exclus. Les articles liés à la santé publique et l’épidémiologie ont été sélectionnés en recherchant les chaînes de caractère « Data mining » ou « Machine learning » dans les titres, les résumés, les mots-clés des auteurs et les termes Mesh. J’ai choisi de cibler les termes génériques « Data mining » et « Machine learning » plutôt qu'un ensemble de techniques, d'algorithmes ou de méthodes spécifiques car les noms des algorithmes utilisés ne sont pas systématiquement cités dans les champs de recherche de Pubmed.

Le titre, le résumé, les mots clés et les termes Mesh de chacun des articles ainsi sélectionnés ont été extraits pour constituer le corpus qui a ensuite été analysé à l'aide des modèles thématiques. Les métadonnées (année de publication, auteurs, PMID et revue de publication) ont été également extraites afin de les inclure dans l’analyse.

2.3.2. Analyse bibliométrique

La première étape de cette analyse consistait en la description du corpus *via* des statistiques descriptives des articles inclus (nombre total d'articles, années de publication et description des revues). La popularité des méthodes DMML pour une année donnée a été évaluée en divisant le nombre d'articles du corpus par le nombre total d'enregistrements dans Pubmed ayant les mêmes caractéristiques (c'est-à-dire : articles, non dupliqués, avec un résumé disponible, sur l'homme, en anglais) publiés l'année en question.

2.3.2.1. Modélisation par l'allocation de Dirichlet latent

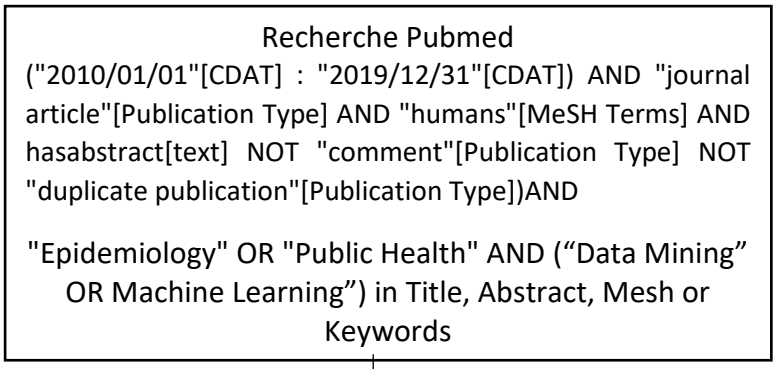
Par la suite, un modèle thématique par allocation de Dirichlet latent (LDA), développé par Blei et al. a été utilisé afin d'identifier les applications les plus prégnantes des méthodes DMML dans le domaine de la santé publique et l'épidémiologie (84). Cette approche a été appliquée à l'analyse bibliométrique de la littérature dans diverses disciplines scientifiques.

L'approche LDA est basée sur un modèle bayésien hiérarchique qui est utilisé pour décomposer une collection de documents en thèmes saillants. Son principe repose sur une assomption majeure : chaque document inclut plusieurs thèmes avec une probabilité d'appartenance différente et les mots qui apparaissent dans ce document reflètent l'ensemble particulier des thèmes qu'il aborde (85). En d'autres termes, chaque document (dans notre cas l'ensemble des titres, résumés, mots clés, termes Mesh) est une distribution multinomiale sur N thèmes latents. Chaque thème a une distribution spécifique de mots qui ont tendance à apparaitre ensemble et dont l'ordre d'apparition n'est pas pris en considération. La modélisation LDA suit un processus d'apprentissage non supervisé, la structure du thème latent est déduite à partir de la probabilité de cooccurrence des mots dans les documents observés. Le processus de modélisation est résumé dans la Figure 1.

Ce modèle permet d'obtenir deux distributions de probabilités postérieures :

i) La distribution des mots par thème P (mot|thème), est interprétée comme le degré d'importance de chaque mot dans un thème. Les mots ayant les plus grandes probabilités d'appartenance à un thème donnent une assez bonne description du thème grâce à leur combinaison.
ii) La distribution des thèmes par document P (thème|document).

Figure 1: Processus du modèle LDA



Extraction des données

Document= PMID, Titre, Résumé, Mots clés, Termes Mesh, Auteur, année de publication

Création du corpus prétraitement des données

Corpus : composé de X documents (résumé, mots clés, Titre, Termes Mesh) + Méta données (dates, revues, PMID)

Application du modèle LDA

Obtention de 30 thèmes Latents

Etape de post-traitement

P(mot|thème)

	Thème 1	Thème 2	Thème 3	….	Thème 30
genetic	0,0137	5,43 e-05	0,005	…	1 e-05
gene	0,0327	5,43 e-05	0,04	…	5,43 e-05
disease	0,0239	0,0078	0,002	…	0,001
genome	0,0064	5,43 e-05	0,1	…	2,00 e-04
……	…..	…..	….	…	…..
somme	Σ =1	Σ =1	Σ =1	Σ =1	Σ =1

P(thème|résumé)

	Résumé 1	Résumé 2	….	Résumé X
Thème 1	0,8	0,25	….	0,0001
Thème 2	0,18	0,7	…	0,15
Thème 3	0,00001	0,00001	…	0,52
Autre thème	…..	……	….	….
Thème 30	0,0002	0,002	….	0,0005
Somme	Σ =1	Σ =1	Σ =1	Σ =1

Interprétation des 30 thèmes latents

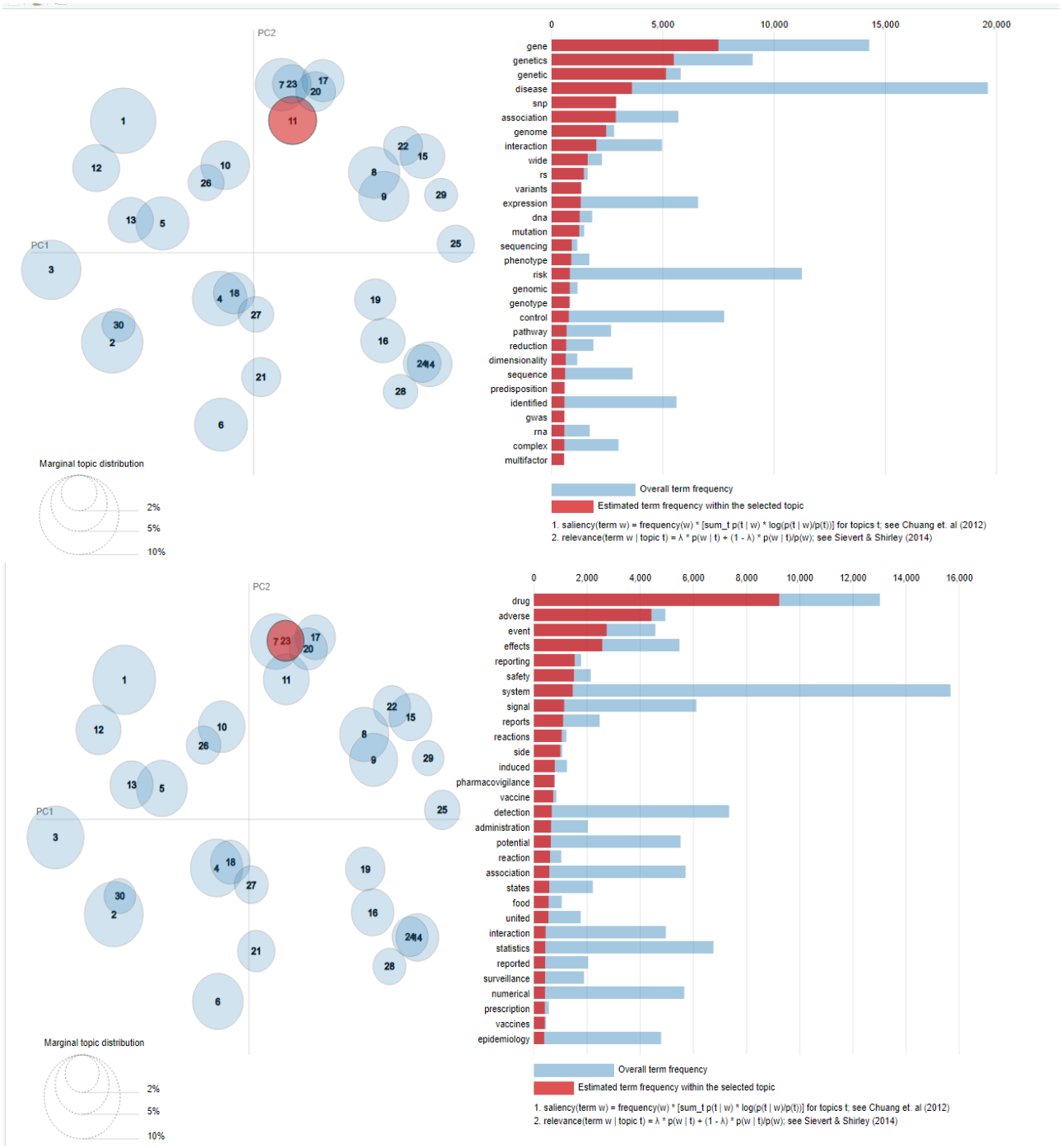
En pratique, la modélisation thématique a débuté par une étape de prétraitement standardisée, qui a permis de nettoyer le corpus. Celle-ci a consisté tout d'abord en la suppression des mots sans valeur analytique (e.g « ceci », « il », « mais »), des mots qui apparaissaient couramment dans les résumés de recherche (e.g « contexte », « objectif », « méthode », « résultat », « conclusion »), des termes apparaissant moins de 20 fois dans le corpus ou apparaissant dans plus de la moitié des documents, des nombres ainsi que de la ponctuation. Les majuscules ont été transformées en minuscules et les mots pluriels ont été convertis au singulier. Les mots restants ont été utilisés pour construire une matrice termes-documents.

Ensuite, le modèle LDA a été appliqué à cette matrice de mots en utilisant la librairie« textmineR » du logiciel R. Les paramètres du modèle ont été spécifiés pour découvrir 30 thèmes avec une grande capacité d'interprétation (hyperparamètres de Dirichlet : alpha 0,1 ; bêta 0,01) en utilisant un échantillonneur Gibbs réduit réglé pour fonctionner sur 5000 itérations (86). La distribution des mots par thème (P(mot|thème)) a été utilisé pour définir le contenu des documents du corpus. En effet, la combinaison des mots ayant les plus fortes probabilités d'appartenance à un thème donnent une assez bonne description du contenu du thème. La distribution des 30 thèmes (P (thème|document)) a été utilisée pour identifier les thèmes majeurs pour chaque article. Conformément aux recommandations d'Antons et al., j'ai considéré qu'un article abordait effectivement un thème si la charge de ce thème sur ce document était supérieure à 0,10 (87).

2.3.2.2. Description et interprétation des thèmes

Les résultats du modèle LDA ont été explorés à l'aide d'un système de visualisation interactif basé sur le web et construit avec la librarie « LDAvis » (88), où les thèmes sont positionnés sur un diagramme en composantes principales basé sur leur relation sémantique (Figure 2). Un histogramme permet de sélectionner les 30 termes les plus pertinents en les sélectionnant sur la fréquence de ces mots à l'intérieur du thème ainsi que leur fréquence sur l'ensemble du corpus. Par ailleurs, les articles représentatifs du thème ont été identifiés en sélectionnant ceux ayant ce thème comme thème principal (i.e une probabilité d'appartenance du thème à l'article supérieure à 0,5). Ces articles ont été utilisés pour illustrer la partie résultats/discussion. Ces articles, seront référencés par leur nom et année de publication, dans le but de les différencier de ceux utilisés dans l'ensemble de la thèse. Sur ces bases, un comité d'adjudication composé de différents chercheurs (en santé publique, médecine générale, psychiatrie, épidémiologie, recherche clinique, bio-informatique et neuroimagerie) a interprété le contenu des thèmes dans une approche itérative et collaborative.

Figure 2: Illustration du système d'illustration interactif LDAvis



Légende : A gauche, diagramme de composantes principales représentant la relation sémantique entre les thèmes. A droite, histogramme des 30 mots les plus pertinents pour chaque thème.

2.4. Evolution temporelle de la popularité des thèmes

L'évolution du nombre d'articles associés à chacun des thèmes a tout d'abord été décrite graphiquement. Dans cette analyse sont considérés les différents thèmes identifiés sur chaque article (la proportion d'un thème sur un article pouvant par définition varier de 0 à 1). Une analyse de tendance a été effectuée pour identifier les thèmes « chauds » dont la proportion sur le corpus a globalement augmenté avec les années, à l'aide de tests de tendance de Cochran-Armitage. Une valeur de p inférieure à l'alpha corrigé par Bonferroni, 0,05/30 = 0,0017, a été considérée comme statistiquement significative.

Une description graphique de la dynamique temporelle du champ de l'épidémiologie a été secondairement effectuée en s'intéressant cette fois-ci uniquement au thème prépondérant sur chaque article (un seul article par thème est retenu, celui ayant la proportion la plus forte sur l'article).

2.5. Résultats/discussion

2.5.1. Description des revues et publications

La requête PubMed a permis d'identifier 12177 articles (Tableau 2).

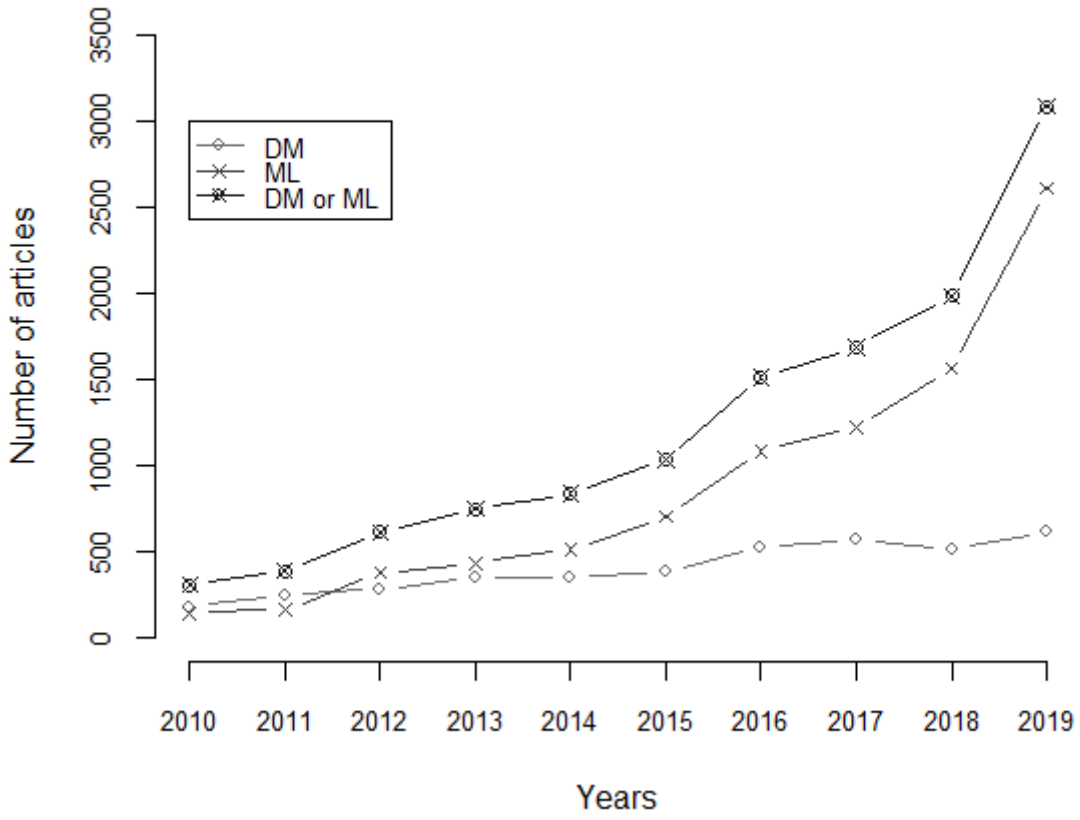
Tableau 2: Description des algorithmes testés et résultats obtenus

	Algorithme testé	Nb articles
# 1	("2010/01/01"[CDAT] : "2019/12/31"[CDAT]) AND "journal article"[Publication Type] AND "humans"[MeSH Terms] AND hasabstract[text] NOT "comment"[Publication Type] NOT "duplicate publication"[Publication Type])AND "Epidemiology" OR "Public Health" in Title, Abstract, Mesh or Keywords [	209 995
# 2	# 1 AND "machine learning" in Title, Abstract, Mesh or Keywords	8 781
# 3	# 1 AND "data mining" in Title, Abstract, Mesh or Keywords	3 978
# 4	# 2 AND # 3	582
# 5	# 2 OR # 3	12 177

Sur les 12 177 articles identifiés de 2010 à 2019, 3978 articles ont été sélectionnés *via* la requête "Data Mining" et 8781 *via* la requête « Machine Learning ». La Figure 3 représente la croissance du nombre de publications liées au DMML dans PubMed. Ce domaine a connu une croissance continue au cours de la dernière décennie, passant de 301 articles en 2010 à 3083 en 2019.

Par rapport au nombre total d'entrées indexées à l'épidémiologie ou à la santé publique dans PubMed, la proportion d'entrées liées au DMML est passée de 0,14 % à 0,97 %.

Figure 3: Nombre annuel d'articles sélectionnés entre 2010 à 2019.



Légende : « DM » entrées PubMed sélectionnées via la requête "Data Mining", « ML » entrées PubMed sélectionnées via la requête "Machine learning", DMML entrées PubMed sélectionnées *via* les requêtes "Data Mining" ET/OU "Machine learning ».

Au total, 1848 revues ont publié ces 12177 articles. Cependant, 61,8% de ces revues (1142/1848) ont publié un ou deux articles au cours de la décennie et seulement 9,9% (183/1848) ont publié dix articles ou plus. Parmi ces dernières revues, quelques revues consacrées à l'épidémiologie ou à la santé publique ont publiés des articles de DMML : « Am J Epidemiol » (10 articles), « BMC Public Health » (13 articles), « Genet Epidemiol » (18 articles), « J Clin Epidemiol » (20 articles), « International Journal of Environmental Research and Public Health » (32 articles) et « Pharmacoepidemiol Drug Saf » (33 articles). Le Tableau 3 classe les 20 revues les plus prolifiques c'est à dire ayant publié le plus d'articles de notre corpus. PLOS One avec 727 articles (5,97% des 12 177 articles) a été de loin la revue la plus prolifique, suivie de deux revues d'informatique bio/médicale fournissant chacune près de 365 articles.

Dans l'ensemble, les revues prédominantes de cette liste sont les revues spécialisées en « Informatique médicale » ou « Bioinformatique » (14/20) tel que le défini Pubmed dans sa page « broad subject terms ».

Tableau 3: Classement des 20 revues ayant le plus publiées d'articles par ordre croissant.

Revue	Nombre d'articles‡	broad subject terms
PLoS One	727	Medicine/Science
Stud Health Technol Inform	364	Health Services Research/Medical Informatics/Technology
Conf Proc IEEE Eng Med Biol Soc	363	Biomedical Engineering
Sci Rep	249	Natural and clinical sciences
J Biomed Inform	248	Medical Informatics
BMC Bioinformatics	220	Computational Biology
Comput Methods Programs Biomed	210	Medical Informatics
J Am Med Inform Assoc	210	Medical Informatics
AMIA Annu Symp Proc	164	Medical Informatics
Comput Biol Med	160	Biology/Medical Informatics/Medicine
IEEE J Biomed Health Inform	138	Biomedical Engineering/Medical Informatics
J Med Syst	137	Medical Informatics
Artif Intell Med	132	Medical Informatics
Comput Math Methods Med	121	Medical Informatics
Sensors (Basel)	116	Biotechnology/Technology
BMC Med Inform Decis Mak	104	Biomedical Engineering
Neuroimage	103	Diagnostic Imaging
IEEE Trans Biomed Eng	99	Biomedical Engineering
Bioinformatics	91	Bioinformatics
Int J Med Inform	91	Medical Informatics

Légende : ǂ Nombre de publications du Corpus par cette revue

2.5.2. Description des domaines d'application

Le Tableau 4 présente les thèmes construits par le modèle LDA. A chaque article a été attribué au moins un thème, le nombre médian de thèmes identifiés par article étant de 3 (intervalle interquartile [2-4]). Comme attendu, les thèmes étaient très divers et reflétaient la variabilité inter et intra résumé des articles : i) des domaines très généraux de la médecine à la biologie (e.g neurophysiologie), ii) des champs d'applications (e.g prédiction de pronostics clinques), iii) des types de données (e.g dossiers médicaux électroniques, images radiographiques), iv) les méthodes DMML (e.g classification, traitement automatique du langage…). Les deux derniers thèmes identifiés par le modèle LDA regroupaient du vocabulaire courant des résumés (e.g méthode, approche, problème, ou recherche, information) – mais peu informatifs.

Le thème le plus fréquemment rencontré était un thème de méthodes DMML et regroupait du vocabulaire lié aux tâches de classification (Thème 30: 22,65% des articles). Parmi les autres thèmes caractérisant les méthodes DMML utilisées, on trouve logiquement la mesure de performances des classificateurs (Thème 23) puis par ordre décroissant de fréquences, les modèles de régression (Thème 24), les réseaux complexes (Thème 17: e.g réseaux bayésiens, les réseaux booléens probabilistes, les modèles de Markov cachés), les algorithmes de clusterisation (Thème 20) et les réseaux de neurones (Thème 15).

Les algorithmes de fouille de données textuelles ont été regroupés ensemble (Thème 10 : Traitement automatique du langage, (Pradhan et al. 2015; Meystre et al. 2010)), l'application typique étant l'extraction automatique de données issues de documents cliniques électroniques (Thème 8 e.g (Dixon et al. 2017; Parke II, Lum, and Rich 2017)). Ceci à des fins multiples : estimation de l'incidence et la prévalence des maladies, de l'efficacité des traitements, suivi des événements indésirables, management de l'utilisation des services de soins, évaluation du respect des « bonnes pratiques » et des directives cliniques.

De nombreux articles analysent des données « omiques » : génomique (Thème 9), transcriptomique, (Thème 11) protéomique et métabolomique (Thème 5) générées par les technologies issues de la biologie cellulaire, biochimie, et chimie. Outre la sélection des SNPs (Smedley et al. 2016) dans les études génomiques, les algorithmes DMML ont été particulièrement utilisés pour explorer les interactions gène-gène (Visweswaran, Wong, and Barmada 2009; Peng, Tang, and Xie 2018) et les modifications épigénétiques (He et al. 2015) qui peuvent moduler le risque associé de maladie. En ce qui concerne le « profilage de l'expression des gènes », une application typique était l'identification de marqueurs cancéreux micro ARN (Dong and Xu 2019) long ARN non codant (Wang et al. 2018) ou exons d'ARNm (Schramm et al. 2012) prédicteurs du pronostic du patient ou de la réponse aux

traitements. De même, les méthodes DMML ont été appliquées à l'analyse des « big data » protéiques et métaboliques (Thème 5) pour la recherche de biomarqueurs potentiels diagnostiques et pronostiques (Thème 5, eg (Bjornevik et al. 2019; Chen et al. 2017; Htun et al. 2017)). Le thème « biologie et chimie computationnelle » (Thème 1) concerne l'utilisation des méthodes DMML pour la prédiction in silico des propriétés des données biologiques et chimiques : prédiction des propriétés structurelles, fonctionnelles ou interactionnelles des protéines et des petites molécules. Les applications étaient très diverses, un exemple en étant la prédiction de toxicité dans le cadre de l'évaluation des risques environnementaux sur la santé humaine (Ge et al. 2015; McPhail et al. 2012; Chen et al. 2014).

Les types de données utilisés comprenaient bien d'autres sources tels que les données de capteurs mobiles et portables (Thème 16 e.g (Dobkin and Dorsch 2011; Kerr et al. 2016; Thralls et al. 2019)), d'imagerie radiographique médicale (Thème 22, e.g tomodensitométrie (Choi et al. 2018)), d'imagerie par résonance magnétique (Thème 26, 6 16, 18), ou les bio signaux (Thème 26 e.g électrocardiographiques (Hermans et al. 2018), encéphalographies (Zhou, Wu, and Zeng 2015)). A titre d'exemple, des données de mesures de l'activité physique issues d'accéléromètres sont de plus en plus intégrées dans les études cliniques voire dans les cohortes en population générale en plus des questions sur l'activité physique. En ce qui concerne l'imagerie, les techniques d'apprentissage automatique ont été particulièrement utilisées pour améliorer le traitement de l'image (Thème 18 (Gloger et al. 2018; Habes et al. 2013)) dans un objectif diagnostic.

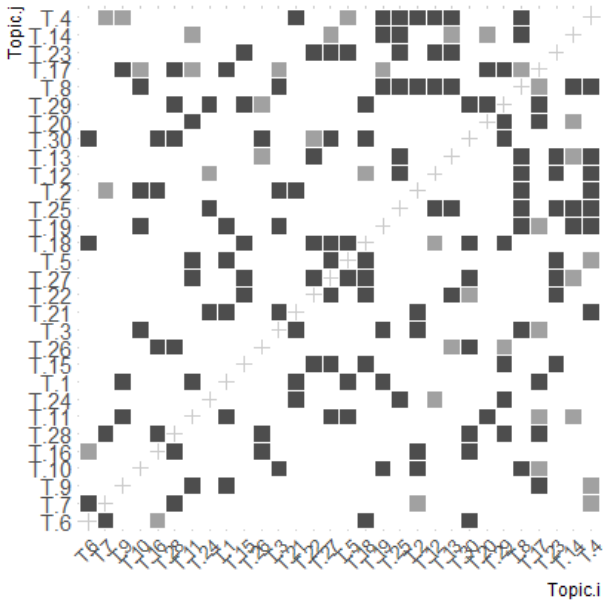
Parmi les pathologies les plus étudiées dans ce corpus, émergent le cancer (Thème 3 notamment cancer du sein, de la prostate, du poumon, carcinomes, glioblastomes), le diabète et les pathologies cardiovasculaires (Thème 2), les troubles neurologiques (Thème 6 notamment maladie d'Alzheimer et autres démences, la maladie de Parkinson, les troubles cognitifs de la personne âgée, la sclérose en plaques) et les maladies psychiatriques (Thème 7 notamment schizophrénie, troubles du spectre autistique, ADHD, psychoses).

Comme le souligne la Figure 4, les méthodes DMML ont été impliquées dans des articles où il est question d'affiner le diagnostic de ces pathologies via des données d'imagerie, et par la recherche de biomarqueurs issus du profilage d'expression des tissus, et des « big data » (bio)chimiques (eg spectrométrie de masse):

- Imagerie pour les troubles neurologiques et psychiatriques (thèmes 28, 7 et 6) et en oncologie,
- Données d'expression génique des tissus pour le typage des tumeurs

- Biomarqueurs chimiques pour des pathologies très variées, allant de cancers, au diabète jusqu'aux maladies psychiatriques.

Figure 4: Matrice de cooccurrence de thèmes sur les articles



Légende : En gris sont représentées les cooccurrences statistiquement significativement plus fréquentes que celles attendues par les distributions

Un autre pan du corpus s'intéressait aux outils d'aide à la décision clinique (Thèmes 25,14 et 12). En effet, avec l'extension des registres médicaux électroniques, les méthodes DMML sont devenues populaires pour modéliser l'état de santé des patients, et leur pronostic. Les nombreuses applications retrouvées dans le corpus incluaient la prédiction du devenir du patient en termes de récidives ou survie, de durée d'hospitalisation, et du risque de ré-hospitalisation… (Thème 25 (Chiew et al. 2019; Nanayakkara et al. 2018; Morgan et al. 2019)) dans divers domaines de la médecine, dont l'oncologie, les maladies cardiovasculaires, et la santé mentale. Ainsi, des algorithmes DMML ont permis de fabriquer des outils capables de dériver des règles individualisées à chaque hôpital et actualisables dans le temps au fur et à mesure que de nouvelles données apparaissent ; le tout en un processus unique et généralisable (Morgan et al. 2019). Les méthodes DMML ont aussi été utilisées pour modéliser la réponse au traitement (Thème 14 (Chekroud et al. 2016; Waljee et al. 2018; Westborg and Rosso 2018)) afin d'identifier les patients les plus susceptibles de bénéficier de chaque thérapie notamment en utilisant des données d'essais cliniques contrôlés. Dans le domaine de la chirurgie, (Thème 12) les méthodes DMML ont été appliquées dans deux domaines très différents : la robotique et l'épidémiologie clinique. Les modélisations du risque opératoire, deviennent courantes pour

identifier les candidats à risque de complications postopératoires, et appréhender les facteurs de risque (Xie et al. 2018; Lötsch, Ultsch, and Kalso 2017; Wojciuk et al. 2015).

D'une manière générale l'utilisation de l'apprentissage automatique pour la médecine personnalisée s'est avéré être un champ d'application conséquent. Leur capacité à exploiter un nombre très important de variables en fait des candidats populaires pour l'individualisation et la reconnaissance des différences chez un patient qui peuvent le rendre phénotypiquement différent de ses pairs. Pour autant certains thèmes identifiés par le modèle LDA étaient en dehors du champ de la recherche clinique.

En effet, un nombre important des articles du corpus concernaient la surveillance épidémiologique. Les systèmes de surveillance retrouvés (Thème 21) incluaient notamment le suivi des épidémies et des conditions environnementales. Les approches DMML y sont souvent utilisées pour détecter l'apparition de signaux dans l'espace et dans le temps. Par exemple, pour produire des systèmes d'alerte précoce des épidémies saisonnières de grippe à partir des données de Twitter (Aslam et al. 2014) pour modéliser la distribution spatiale de certains vecteurs de maladies (Hernández et al. 2013), améliorer le système de surveillance du virus VIH (Oliveira et al. 2017), analyser les risques d'accident en temps réel (Wang et al. 2019), les tendances spatio-temporelles des micro particules (Meng et al. 2018). Par ailleurs, de nombreuses études ont utilisé les algorithmes de fouille de données numériques et textuelles pour détecter les événements indésirables liés aux médicaments (Thème 19 pharmacovigilance) *via* l'exploration des données des systèmes de notification des événements indésirables (e.g (Moro et al. 2016; Kadoyama et al. 2011), de bases de données pharmaceutiques, de la littérature biomédicale (Gurulingappa et al. 2013), ou des réseaux sociaux (Yang and Yang 2018).

Le thème 4 (e.g (Schoos et al. 2016; Kuhle et al. 2018; Heo and Ryu 2018)), peu informatif, rassemble du vocabulaire lié à la description des populations incluses (sexe, age) ainsi que du vocabulaire typique de l'épidémiologie (cohorte, association, facteurs de risque).

Pour finir, le thème 2 (e.g (Mowery et al. 2017; Mackey et al. 2017; Rho et al. 2019)), inclut les applications épidémiologiques en santé mentale et en santé sociale. Les applications les plus typiques utilisaient les méthodes DMML sur les données issues de réseaux sociaux pour analyser les déterminants de comportements en matière de santé. Les troubles les plus explorés étaient la consommation de substances psychoactives, les comportements suicidaires, les troubles anxio-dépressifs ainsi que ceux liés à des traumatismes.

Tableau 4: Description des thèmes obtenus suite à l'application du modèle LDA sur le corpus de documents

N	Nom du thème	Nombre d'articles représentatifs de ce topic	Pourcentage d'articles représentatif par rapport à tout le corpus	Mots les plus fréquents‡	Mots les plus spécifiques du thème*
1	Biologie et chimie computationelle	1212	9,95	protein, prediction, metabolism, sequence, drug	chemicals, qsar, motifs, docking, epitopes
2	Epidémiologie sociale et en santé mentale	1002	8,23	health, social, disorder, media, psychology	suicide, tewwets, twiter, questionnaire, sentiments
3	Thème général	2651	21,77	research, information, system, tool, clinical	papers, publication, gaps, ethical, legal
4	Description des populations incluses dans les études (épidémiologie)	1252	10,28	age(d), female, adult, child, years	infant, asthma, pregnancy, birth,newborn
5	Bio marqueurs (Proétome et métabolome)	967	7,94	blood, biomarker, diagnosis, mass, sample	spectrometry, retinal, urine, metabolomics, glaucoma
6	Troubles neurologiques	717	5,89	disease, ad (Alzheimer's disease), cognitive, diagnosis, brain	Alzheimer, dementia, mci, parkinson, atrophy
7	Neuropsychiatrie	894	7,34	brain, mri, disorder, functional, control	schizophrenia, autism, asd, adhd, mdd
8	Dossiers de santé électroniques	1786	14,67	patient, electronic, care, health, records	administrative, icd, emr, insurance, medicare
9	Epidémiologie Génétique	1308	10,74	gene, genetic, disease, SNP, association	SNP predisposition, GWAS, multifactor, mdr (multifactor dimensionality reduction)
10	Traitement automatique du langage	1104	9,07	clinical, text, natural, system, language	nlp (natural langage processing), vocabulary, summary, narrative, note
11	Profilage de l'expression génique	1403	11,52	gene, cell, expression, cancer, genetics	upregulated, bci (Breast Cancer Index), cxcl, mesenchymal, qPCR
12	Chirurgie	632	5,19	patient, surgery/surgical, pain, aged, complication	postoperative, dental, spinal, wound, spine
13	Diabète et maladies cardiovasculaires	899	7,38	diabetes, disease, heart, patient, coronary, cardiovascular	coronary, artery, mellitus, myocardial, infarction
14	Prédiction de la réponse au traitement	933	7,66	patient, treatment, therapy, clinical, drug	radiotherapy, rheumatoid, responders, systemic lupus erythematosus, placebo
15	réseaux de neurones	744	6,11	learning, deep, ANN convolution, CNN	
16	Mesure du mouvement	913	7,5	activity, sensor, system, monitoring, time	sensor, gait, fall, walking, smartphone
17	Réseaux complexes e.g bayésiens, booléens...	1561	12,82	network, bayesian, disease, approach, knowledge	Bayesian networks, abstraction, vice versa, logic, unveil
18	Traitement de l'image (IRMs…)	1697	13,94	image/imaging, computer, segmentation, automated, assisted	segmentation, dice, pixel, patche, contour

N	Nom du thème	Nombre d'articles représentatifs de ce topic	Pourcentage d'articles représentatif par rapport à tout le corpus	Mots les plus fréquents ǂ	Mots les plus spécifiques du thème*
19	pharmacovigilance	676	5,55	drug, adverse, event, effects, reporting	pharmacovigilence, adr (adverse drug reaction), adverse events, FDA, vaers (vaccine adverse events reporting system)
20	Méthodes de clusterisation	792	6,5	cluster, clustering, algorithm, unsupervised, pattern	cluster, tcm, tongue, biclustering, acupuncture
21	Surveillance de la santé publique (infection environnementale)	892	7,33	disease, HIV, epidemiology, surveillance, infection	crash, microbiology, traffic, accidents, dengue
22	Imagerie médicale radiographique	695	5,71	ct (Computed Tomography), imaging, diagnostic, tomography, liver	ct (Computed Tomography), hcc (Hepatocellular carcinoma), fibrosis, nodule, cirrhosis
23	Mesure de performance des modèles	2028	16,66	algorithm, curve, sensitivity, ROC, AUC	PPV, NPV, cutoffs, receiver, negatives
24	Modèles de régression	1841	15,12	regression, prediction, logistic, statistical, variables	RMSE, imputation, squared, MARS, covariate
25	Prédiction du risque clinique	1401	11,51	risk, patient, prediction, mortality, outcome	sepsis,intensive care unit, readmission, Acute kidney injury, glasgow
26	Bio signaux eg EEG, *ECG*, and *PPG*	871	7,15	signal, EEG, sleep, detection, feature	seizure, hearing, epileptic, apnea, fibrillation
27	Cancer	1301	10,68	cancer, breast, tumor, neoplasms, patient	glioblastoma, glioma, adc, gleason, idh
28	Neurophysiologie	699	5,74	physiology, brain, eeg, computer, visual	bci (brain–computer interface), decoding, evoked, erp (event-related potential), stimulus
29	Thème général	2453	20,15	algorithm, method, learning, proposed, approach	manifold, sparsity, subspace, norm, compressed
30	Méthodes de classification	2758	22,65	SVM, feature, classification, classifier, accuracy	rbf (Radial Basis Function), multiclass, rfe (Recursive Feature Elimination), oversampling, smote (synthetic minority oversampling technique)

Légende: ǂ Mots avec les occurrences les plus élevées. * Mots dont le taux (occurrences dans le thème/ occurrences dans tout le corpus)

2.5.3. Dynamique temporelle des différents domaines

En termes de nombre d'articles, tous les thèmes sans exception montraient globalement une augmentation au cours du temps (Figure 5 ; Figure 6 ; Figure 7). Les raisons de cette augmentation relèvent non seulement de l'augmentation du nombre d'études utilisant des méthodes DMML complexes (ANN, SVM, RF, SGD) mais aussi de l'expansion du nombre total d'articles enregistrés dans PubMed chaque année. Cette augmentation peut également être due au référencement plus fréquent avec les expressions « Machine Learning » ou « Data Mining » à type de technique identique par effet de mode. En effet, comme soulignée dans l'introduction de ce chapitre, les méthodes de régression pénalisée ou les arbres de régression et de classification étaient classiquement considérés du domaine de la statistique alors que sur la dernière décennie ces mêmes modèles ont été fréquemment référencés par les auteurs comme du « Machine Learning ».

Figure 5 : Evolution du nombre d'articles publiés par année pour chacun des thèmes (thème 1 à 10)

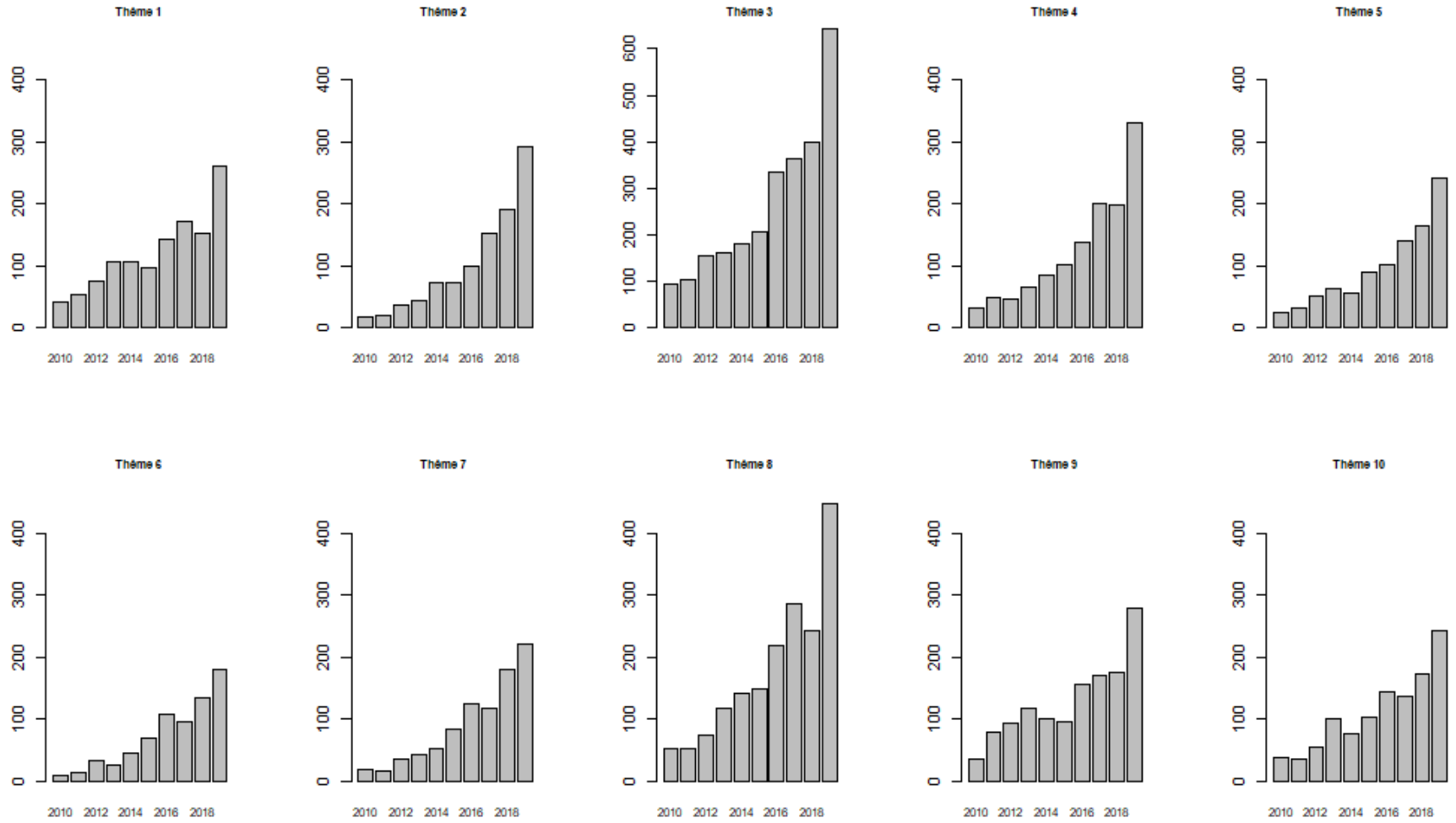


Figure 6 : Evolution du nombre d'articles publiés par année pour chacun des thèmes (thème 11 à 20)

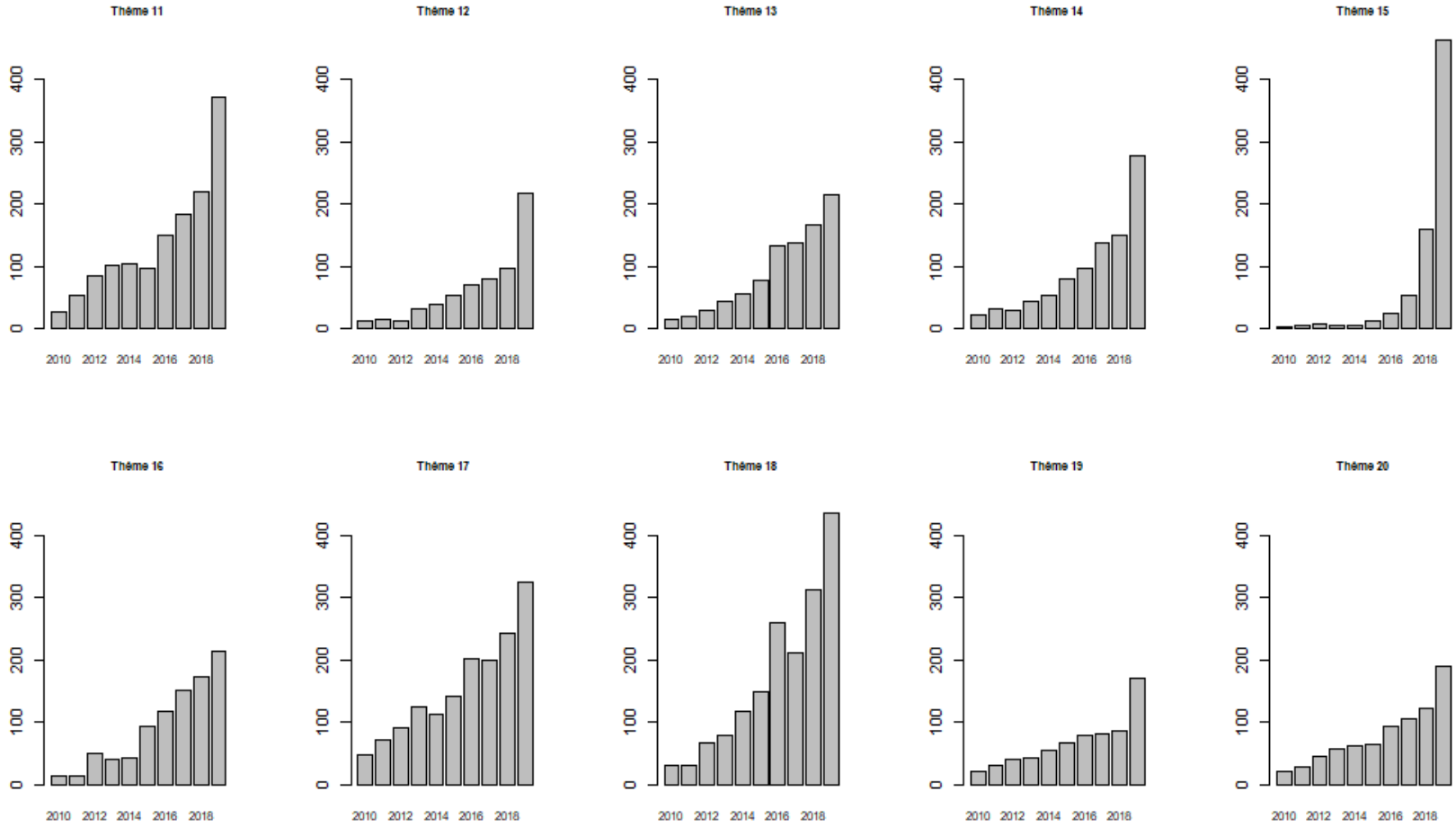
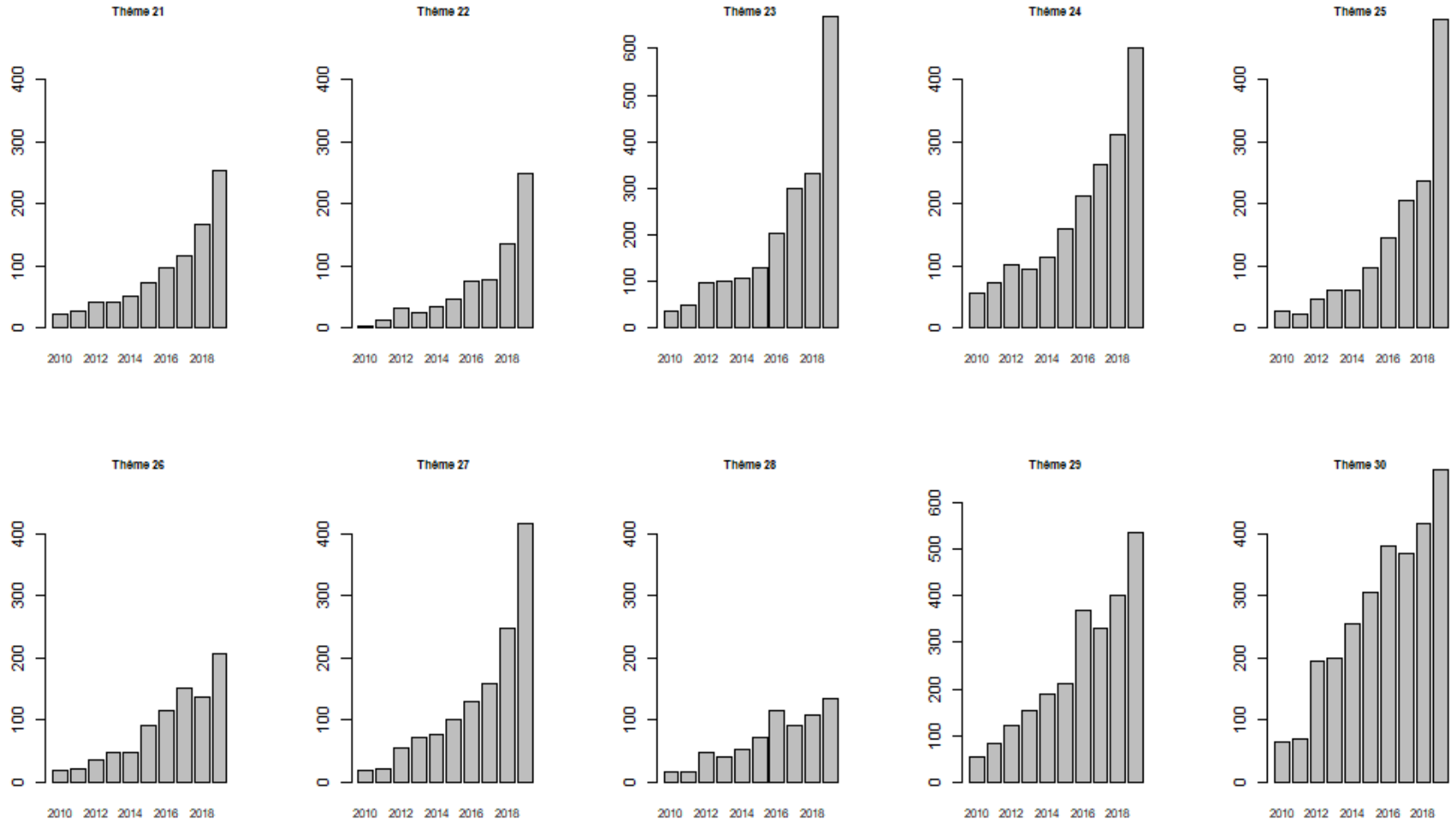
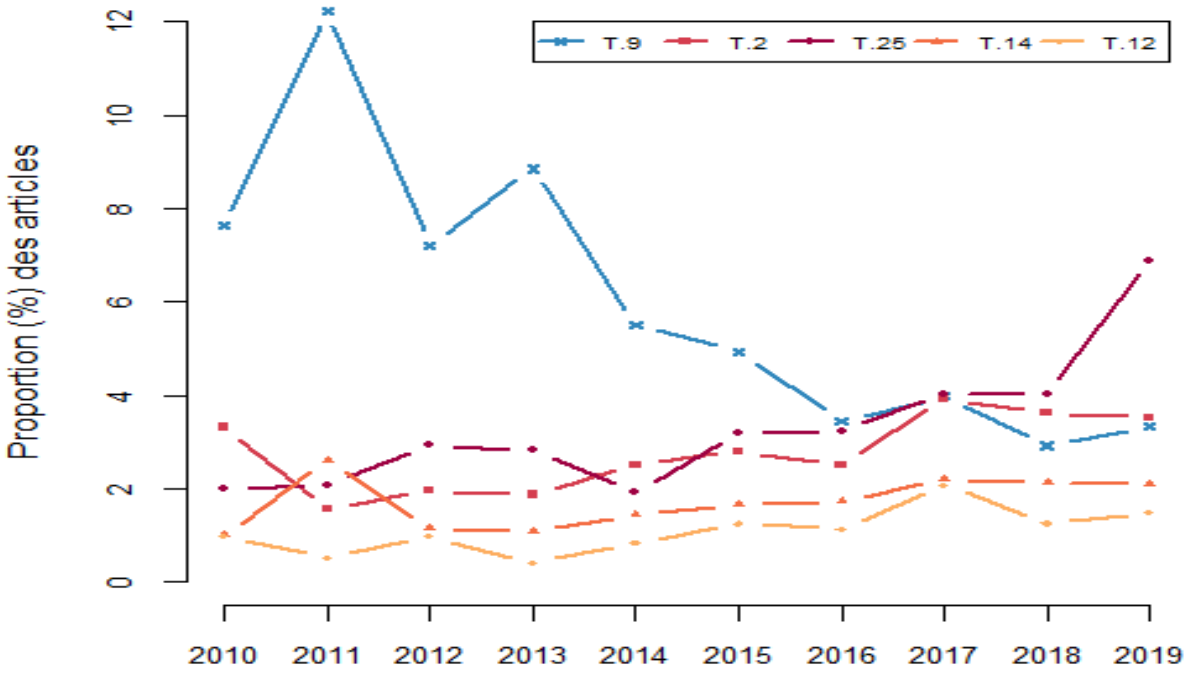


Figure 7 : Evolution du nombre d'articles publiés par année pour chacun des thèmes (thème 21 à 30)



Ceci étant, relativement à l'ensemble du corpus, certains thèmes ont montré une augmentation plus forte que d'autres. Ces sujets « chauds », portés par une proportion croissante d'articles au cours des années étaient : Les réseaux de neurones (Thème 15 test de tendance de Cochran-Armitage p<1e-5), la mesure de la performance des modèles (Thème 23 p<1e-5), le traitement de l'image (Thème 18 p <1e-05), l'imagerie médicale radiographique (Thème 22 p<1e-5), le cancer (Thème 27 p<1e-5), l'épidémiologie sociale et en santé mentale (Thème 2 p<1e-5), recherche clinique (prédiction du risque, de la réponse au traitement et de la chirurgie (Thème 25 p<1e-5, Thème 14 p<1e-3et Thème 12 p<1e-5). Par contre, le thème de l'épidémiologie génétique (Thème 9) qui s'était déjà en partie approprié les approches DMML au début de la décennie a relativement moins augmenté ces dernières années que les branches de l'épidémiologie relevant de la recherche clinique, de la santé mentale et des sciences sociales (voir Figure 8 qui montre le taux d'articles où ces thèmes sont en 1ère position).

Figure 8: Représentation de l'évolution dans le temps de la proportion des articles sur le corpus où les thèmes d'épidémiologie sont le sujet principal de l'article.



Légende : T.2 : Thème « Epidémiologie sociale et en santé mentale » ; T.9 : Thème « Epidémiologie Génétique » ; T.12 : Thème « Chirurgie » ; T.14 : Thème « Prédiction de la réponse au traitement » ; T.25 : Thème « Prédiction du risque clinique ».

2.5.4. Points forts et limites

Les points forts de cette étude comprenaient l'utilisation de la modélisation thématique LDA suivant un processus d'apprentissage non supervisé. Cet outil de modélisation a permis d'extraire la structure cachée d'une grande collection de résumés d'articles s'intéressant à la santé publique/épidémiologie et aux approches DMML. Le volume de la littérature scientifique à ce sujet étant en augmentation permanente, les méthodes de text mining sont idéales pour identifier les informations pertinentes. Une autre force de cette étude était l'ouverture du critère de recherche : aucune restriction sur le type d'algorithme n'a été effectuée. Ces choix ont permis d'avoir accès à un nombre très important d'articles non biaisés par les limites d'une liste – forcément non exhaustive – de noms d'algorithmes. De plus l'utilisation d'une approche non supervisée pour définir les thèmes a pour avantage de dessiner une cartographie objective et sans hypothèses a priori sur les domaines d'applications.

Néanmoins cette modélisation avait ses limites. D'une part, comme tout modèle probabiliste, la modélisation par LDA est limitée par ses hypothèses (16,19). En effet, les modèles LDA appréhende les textes comme un "sac de mots", c'est-à-dire que les mots sont considérés indépendants les uns des autres et l'ordre des mots est ignoré. Une autre limite relève de la subjectivité potentielle lors de l'interprétation du contenu des thèmes. Pour pallier les biais potentiels d'interprétation, un comité d'adjudication, composé de chercheurs, spécialisés dans différents domaines de la recherche en santé publique et épidémiologie a été constituée. Dans un premier temps, chaque membre du comité d'adjudication a interprété et nommé les thèmes. Par la suite, les noms donnés aux thèmes ont été confrontés et les divergences ont été résolues par consensus. Une ultime vérification de l'interprétabilité des domaines a été effectuée.

Enfin, le corpus analysé a été extrait automatiquement de la base PubMed. Contrairement à une revue systématique, nous n'avons pas interrogé d'autres bases de données de la littérature biomédicale, ni tenté de repérer manuellement les articles manquants. Il est donc évident que ce corpus n'est pas exhaustif, puisqu'il manque les articles de revues non indexées dans PubMed et que l'algorithme de recherche automatique, comme tout algorithme a forcément raté une partie des articles d'intérêt. *A contrario*, une partie des résumés sélectionnés a forcément été inclus à tort par exemple à cause de mots clefs des auteurs mal définis. Comme seule une partie des 12177 résumés sélectionnés a été lue, ces inclusions à tort n'ont pas été écartées de l'analyse. Néanmoins, l'objectif de cette revue était de repérer les principaux domaines d'applications de la fouille de données en santé publique ; il est peu probable que la non exhaustivité du corpus ou les erreurs d'inclusions aient biaisé de façon suffisamment marqué le modèle LDA pour cacher des domaines d'applications.

Faute de revues systématiques ou bibliométriques sur le sujet, il est intéressant de comparer ces résultats avec ceux de revues narratives récentes. Dans sa revue « What is Machine Learning? A Primer for the Epidemiologist » publié dans American Journal of epidemiology, (82) . Bi et al. ont listé sept types d'applications épidémiologiques des techniques d'apprentissage automatique publiés dans la littérature : i) Le diagnostic des maladies, ii) les modèles prédictifs pronostic et autres outils d'aide à la décision clinique, iii) les études d'association génomiques, iv) la fouille de données textuelles (en particulier des fichiers de santé électroniques) , v) la prédiction et prévision des maladies infectieuses, vi) les applications géospatiales, et viii) l'inférence causale. Les résultats du modèle LDA obtenu dans ce chapitre sont cohérents avec la structuration de cette revue narrative si ce n'est que le thème « surveillance de la santé publique » créé par le modèle LDA regroupait la prévision des maladies infectieuses et les applications géospatiales. Les co-occurrences mises en évidence entre les thèmes LDA sont en accord avec les exemples d'applications qu'ils ont développés à savoir l'utilisation extensive des SVMs pour la stratification des cancers à partir de données radiologiques et d'expression des gènes ; l'intérêt du text mining pour l'utilisation efficace des fichiers de santé électroniques en recherche clinique ou pour la surveillance.

Par contre le modèle LDA n'a pas identifié de thèmes autours de l'inférence causale. Par cette appellation, les auteurs ont fait principalement référence à l'utilisation des algorithmes d'apprentissage automatique dans l'estimation des scores de propension en présence de données à haute dimension. Le modèle LDA, paramétré pour identifier les 30 thèmes les plus prégnants, n'a pas su identifier cette utilisation peu fréquente des modèles DMML. D'autant que par rapport aux approches statistiques ou épidémiologiques classiques, les algorithmes d'apprentissage automatique ont historiquement mis moins l'accent sur la prédiction plutôt que sur l'inférence étiologique.

Au final, si la majorité des domaines d'applications que notre analyse a identifiée, étaient déjà évoqués dans les revues narratives ou articles d'opinion, l'utilisation d'une technique de fouille de donnée a permis de quantifier l'importance de chaque domaine et de mettre en évidence des utilisations moins souvent citées comme la pharmacovigilance ou les utilisations pour l'épidémiologie sociale et en santé mentale.

2.5.5. Conclusion

En conclusion, cette revue bibliométrique de la littérature a confirmé l'expansion de la popularité des méthodes DMML en santé publique et en épidémiologie ; le nombre d'articles publiés a été multiplié d'un facteur 10 sur la dernière décennie. Néanmoins cette expansion est loin de toucher tous les domaines et les revues qui ont publié régulièrement des articles alliant santé publique et DMML ont

tendance à être des revues à orientation informatique. Les grandes revues épidémiologiques et cliniques ne se sont intéressés à l'apport de ces méthodes que sporadiquement. Les applications des méthodes DMML se sont concentrées sur la recherche de biomarqueurs pour le diagnostic, l'épidémiologie clinique pour prévoir les risques, la progression de la maladie et les résultats des traitements, la surveillance en santé publique et la pharmacovigilance ainsi que les sciences sociales et la santé mentale.

Liste des références permettant de décrire les thèmes de l'analyse bibliométrique

Aslam, Anoshé A, Ming-Hsiang Tsou, Brian H Spitzberg, et al.
2014 The Reliability of Tweets as a Supplementary Method of Seasonal Influenza Surveillance. Journal of Medical Internet Research 16(11). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4260066/, accessed December 24, 2020.

Bjornevik, Kjetil, Zhongli Zhang, Éilis J. O'Reilly, et al.
2019 Prediagnostic Plasma Metabolomics and the Risk of Amyotrophic Lateral Sclerosis. Neurology: 10.1212/WNL.0000000000007401.

Chekroud, Adam Mourad, Ryan Joseph Zotti, Zarrar Shehzad, et al.
2016 Cross-Trial Prediction of Treatment Outcome in Depression: A Machine Learning Approach. The Lancet Psychiatry 3(3): 243–250.

Chen, Huiling, Lufeng Hu, Huaizhong Li, et al.
2017 An Effective Machine Learning Approach for Prognosis of Paraquat Poisoning Patients Using Blood Routine Indexes. Basic & Clinical Pharmacology & Toxicology 120(1): 86–96.

Chen, Yingjie, Feixiong Cheng, Lu Sun, et al.
2014 Computational Models to Predict Endocrine-Disrupting Chemical Binding with Androgen or Oestrogen Receptors. Ecotoxicology and Environmental Safety 110: 280–287.

Chiew, Calvin J., Nan Liu, Takashi Tagami, et al.
2019 Heart Rate Variability Based Machine Learning Models for Risk Prediction of Suspected Sepsis Patients in the Emergency Department. Medicine 98(6). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6380871/, accessed December 23, 2020.

Choi, Kyu Jin, Jong Keon Jang, Seung Soo Lee, et al.
2018 Development and Validation of a Deep Learning System for Staging Liver Fibrosis by Using Contrast Agent–Enhanced CT Images in the Liver. Radiology 289(3): 688–697.

Dixon, Brian E, Guoyu Tao, Jane Wang, et al.
2017 An Integrated Surveillance System to Examine Testing, Services, and Outcomes for Sexually Transmitted Diseases: 5.

Dobkin, Bruce H., and Andrew Dorsch
2011 The Promise of MHealth: Daily Activity Monitoring and Outcome Assessments by Wearable Sensors. Neurorehabilitation and Neural Repair 25(9): 788–798.

Dong, Jingwei, and Mingjun Xu
2019 A 19-miRNA Support Vector Machine Classifier and a 6-miRNA Risk Score System Designed for Ovarian Cancer Patients. Oncology Reports. http://www.spandidos-publications.com/10.3892/or.2019.7108, accessed December 23, 2020.

Ge, Yue, Maribel Bruno, Kathleen Wallace, et al.
2015 Systematic Proteomic Approach to Characterize the Impacts of Chemical Interactions on Protein and Cytotoxicity Responses to Metal Mixture Exposures. Journal of Proteome Research 14(1): 183–192.

Gloger, Oliver, Robin Bülow, Klaus Tönnies, and Henry Völzke
2018 Automatic Gallbladder Segmentation Using Combined 2D and 3D Shape Features to Perform Volumetric Analysis in Native and Secretin-Enhanced MRCP Sequences. Magnetic Resonance Materials in Physics, Biology and Medicine 31(3): 383–397.

Gurulingappa, Harsha, Luca Toldo, Abdul Mateen Rajput, et al.
2013 Automatic Detection of Adverse Events to Predict Drug Label Changes Using Text and Data Mining Techniques. Pharmacoepidemiology and Drug Safety 22(11): 1189–1194.

Habes, Mohamad, Thilo Schiller, Christian Rosenberg, Martin Burchardt, and Wolfgang Hoffmann
2013 Automated Prostate Segmentation in Whole-Body MRI Scans for Epidemiological Studies. Physics in Medicine and Biology 58(17). IOP Publishing: 5899–5915.

He, Jianlin, Ming-an Sun, Zhong Wang, et al.
2015 Characterization and Machine Learning Prediction of Allele-Specific DNA Methylation. Genomics 106(6): 331–339.

Heo, Byeong Mun, and Keun Ho Ryu
2018 Prediction of Prehypertenison and Hypertension Based on Anthropometry, Blood Parameters, and Spirometry. International Journal of Environmental Research and Public Health 15(11): 2571.

Hermans, Ben J M, Job Stoks, Frank C Bennis, et al.
2018 Support Vector Machine-Based Assessment of the T-Wave Morphology Improves Long QT Syndrome Diagnosis. EP Europace 20(suppl_3): iii113–iii119.

Hernández, Jaime, Ignacia Núñez, Antonella Bacigalupo, and Pedro E Cattan
2013 Modeling the Spatial Distribution of Chagas Disease Vectors Using Environmental Variables and People´s Knowledge. International Journal of Health Geographics 12(1): 29.

Htun, Nay M., Dianna J. Magliano, Zhen-Yu Zhang, et al.
2017 Prediction of Acute Coronary Syndromes by Urinary Proteome Analysis. Ingo Ahrens, ed. PLOS ONE 12(3): e0172036.

Kadoyama, Kaori, Akiko Kuwahara, Motohiro Yamamori, et al.
2011 Hypersensitivity Reactions to Anticancer Agents: Data Mining of the Public Version of

the FDA Adverse Event Reporting System, AERS. Journal of Experimental & Clinical Cancer Research: CR 30: 93.

Kerr, Jacqueline, Ruth E. Patterson, Katherine Ellis, et al.
2016 Objective Assessment of Physical Activity: Classifiers for Public Health. Medicine & Science in Sports & Exercise 48(5): 951–957.

Kuhle, Stefan, Bryan Maguire, Hongqun Zhang, et al.
2018 Comparison of Logistic Regression with Machine Learning Methods for the Prediction of Fetal Growth Abnormalities: A Retrospective Cohort Study. BMC Pregnancy and Childbirth 18(1): 333.

Lötsch, J., A. Ultsch, and E. Kalso
2017 Prediction of Persistent Post-Surgery Pain by Preoperative Cold Pain Sensitivity: Biomarker Development with Machine-Learning-Derived Analysis. British Journal of Anaesthesia 119(4): 821–829.

Mackey, Tim K., Janani Kalyanam, Takeo Katsuki, and Gert Lanckriet
2017 Twitter-Based Detection of Illegal Online Sale of Prescription Opioid. American Journal of Public Health 107(12): 1910–1915.

McPhail, Brooks, Yunfeng Tie, Huixiao Hong, et al.
2012 Modeling Chemical Interaction Profiles: I. Spectral Data-Activity Relationship and Structure-Activity Relationship Models for Inhibitors and Non-Inhibitors of Cytochrome P450 CYP3A4 and CYP2D6 Isozymes. Molecules 17(3): 3383–3406.

Meng, Xia, Jenny L. Hand, Bret A. Schichtel, and Yang Liu
2018 Space-Time Trends of PM2.5 Constituents in the Conterminous United States Estimated by a Machine Learning Approach, 2005-2015. Environment International 121(Pt 2): 1137–1147.

Meystre, Stéphane M, Julien Thibault, Shuying Shen, John F Hurdle, and Brett R South
2010 Textractor: A Hybrid System for Medications and Reason for Their Prescription Extraction from Clinical Text Documents. Journal of the American Medical Informatics Association : JAMIA 17(5): 559–562.

Morgan, Daniel J., Bill Bame, Paul Zimand, et al.
2019 Assessment of Machine Learning vs Standard Prediction Rules for Predicting Hospital Readmissions. JAMA Network Open 2(3).
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6484642/, accessed December 23, 2020.

Moro, Pedro L., Emily Jane Woo, Wendy Paul, et al.
2016 Post-Marketing Surveillance of Human Rabies Diploid Cell Vaccine (Imovax) in the

Vaccine Adverse Event Reporting System (VAERS) in the United States, 1990–2015. PLoS Neglected Tropical Diseases 10(7): e0004846.

Mowery, Danielle, Hilary Smith, Tyler Cheney, et al.
2017 Understanding Depressive Symptoms and Psychosocial Stressors on Twitter: A Corpus-Based Study. Journal of Medical Internet Research 19(2): e48.

Nanayakkara, Shane, Sam Fogarty, Michael Tremeer, et al.
2018 Characterising Risk of In-Hospital Mortality Following Cardiac Arrest Using Machine Learning: A Retrospective International Registry Study. Suchi Saria, ed. PLOS Medicine 15(11): e1002709.

Oliveira, Alexandra, Brígida Mónica Faria, A. Rita Gaio, and Luís Paulo Reis
2017 Data Mining in HIV-AIDS Surveillance System : Application to Portuguese Data. Journal of Medical Systems 41(4): 51.

Parke II, D. W., F. Lum, and W. L. Rich
2017 The IRIS® Registry. Der Ophthalmologe 114(1): 1–6.

Peng, Zhe-Ye, Zi-Jun Tang, and Min-Zhu Xie
2018 Research Progress in Machine Learning Methods for Gene-Gene Interaction Detection. Yi Chuan = Hereditas 40(3): 218–226.

Pradhan, Sameer, Noémie Elhadad, Brett R South, et al.
2015 Evaluating the State of the Art in Disorder Recognition and Normalization of the Clinical Narrative. Journal of the American Medical Informatics Association : JAMIA 22(1): 143–154.

Rho, Mi Jung, Jihwan Park, Euihyeon Na, et al.
2019 Types of Problematic Smartphone Use Based on Psychiatric Symptoms. Psychiatry Research 275: 46–52.

Schoos, Ann-Marie Malby, Bo Lund Chawes, Morten Arendt Rasmussen, et al.
2016 Atopic Endotype in Childhood. The Journal of Allergy and Clinical Immunology 137(3): 844-851.e4.

Schramm, A, B Schowe, K Fielitz, et al.
2012 Exon-Level Expression Analyses Identify MYCN and NTRK1 as Major Determinants of Alternative Exon Usage and Robustly Predict Primary Neuroblastoma Outcome. British Journal of Cancer 107(8): 1409–1417.

Smedley, Damian, Max Schubach, Julius O.B. Jacobsen, et al.
2016 A Whole-Genome Analysis Framework for Effective Identification of Pathogenic

Regulatory Variants in Mendelian Disease. American Journal of Human Genetics 99(3): 595–606.

Thralls, Katie J., Suneeta Godbole, Todd M. Manini, et al.
2019 A Comparison of Accelerometry Analysis Methods for Physical Activity in Older Adult Women and Associations with Health Outcomes over Time. Journal of Sports Sciences 37(20): 2309–2317.

Visweswaran, Shyam, An-Kwok Ian Wong, and M. Michael Barmada
2009 A Bayesian Method for Identifying Genetic Interactions. AMIA Annual Symposium Proceedings 2009: 673–677.

Waljee, Akbar K., Boang Liu, Kay Sauder, et al.
2018 Predicting Corticosteroid Free Endoscopic Remission with Vedolizumab in Ulcerative Colitis. Alimentary Pharmacology & Therapeutics 47(6): 763–772.

Wang, Ling, Mohamed Abdel-Aty, Jaeyoung Lee, and Qi Shi
2019 Analysis of Real-Time Crash Risk for Expressway Ramps Using Traffic, Geometric, Trip Generation, and Socio-Demographic Predictors. Accident Analysis & Prevention 122: 378–384.

Wang, Xin, Lei Han, Ling Zhou, Li Wang, and Lan-Mei Zhang
2018 Prediction of Candidate RNA Signatures for Recurrent Ovarian Cancer Prognosis by the Construction of an Integrated Competing Endogenous RNA Network. Oncology Reports. http://www.spandidos-publications.com/10.3892/or.2018.6707, accessed December 23, 2020.

Westborg, Inger, and Aldana Rosso
2018 Risk Factors for Discontinuation of Treatment for Neovascular Age-Related Macular Degeneration. Ophthalmic Epidemiology 25(2). Taylor & Francis: 176–182.

Wojciuk, Bartosz, Marek Myślak, Krzysztof Pabisiak, Kazimierz Ciechanowski, and Stefania Giedrys-Kalemba
2015 Epidemiology of Infections in Kidney Transplant Recipients – Data Miner's Approach. Transplant International 28(6): 729–737.

Xie, Shangchen, Wenjuan Ma, Minxue Shen, et al.
2018 Clinical and Pharmacogenetics Associated with Recovery Time from General Anesthesia. Pharmacogenomics 19(14): 1111–1123.

Yang, Christopher C., and Haodong Yang
2018 Mining Heterogeneous Networks with Topological Features Constructed from Patient-Contributed Content for Pharmacovigilance. Artificial Intelligence in Medicine 90: 42–52.

Zhou, Jing, Xiao-ming Wu, and Wei-jie Zeng
2015 Automatic Detection of Sleep Apnea Based on EEG Detrended Fluctuation Analysis and Support Vector Machine. Journal of Clinical Monitoring and Computing 29(6): 767–772.

3. Contexte et présentation de l'enquête « Processus d'Adolescence ».

Dans la suite de cette thèse, pour faciliter la lecture du manuscrit, le terme « adolescents » définira l'ensemble des sujets quel que soit le sexe (féminin ou masculin). Le terme « filles », définira le groupe d'adolescents de sexe féminin et le terme « garçons », le groupe d'adolescents de sexe masculin.

3.1. Contexte

En santé publique et épidémiologie, l'analyse des facteurs associés au risque de survenue d'un évènement d'intérêt est généralement réalisée à l'aide d'une régression linéaire ou logistique. Un modèle de régression permet d'établir la relation entre une variable numérique ou binaire et des variables explicatives. C'est un modèle simple, pertinent, relativement facile à implémenter et à interpréter en cas d'utilisation pour une base de données de taille modérée (en termes de nombre d'individus et de variables explicatives). Comme pour toute technique de modélisation, la sélection des variables devant figurer dans le modèle est une étape clé. Pour ce faire, il existe plusieurs approches d'inclusion de variables dans le modèle, par exemple à partir d'une revue de la littérature sur le lien entre la variable d'intérêt et les variables explicatives ; à partir des résultats de l'analyse bivariée (avec p-valeur <0.05 ou p<0.2) ou *via* des procédés d'intégration dans un modèle faisant appel à des processus itératif de type « stepwise » (se basant généralement sur un critère d'ajustement du modèle aux données, tel que le critère d'information d'Akaike (AIC) (89) ou le critère d'information bayésien (BIC) (90)).

Cependant, le développement de grands jeux de données en épidémiologie soulève de nombreux problèmes. Des problèmes de robustesse et de stabilité se posent lorsque le nombre de variables explicatives est trop important, de sur-ajustement du modèle aux données, de gestion du risque de première espèce suite à des tests statistiques multiples. L'estimation par le maximum de vraisemblance généralement utilisé dans ce cas, pose des difficultés multiples liées notamment à de fortes inter-corrélations entre variables explicatives, ou encore une complexité de calculs informatiques. Le nombre d'interactions potentielles au second ou troisième degré devient considérable. Les méthodes statistiques qui nécessitent l'utilisation de la significativité statistique deviennent alors difficilement interprétables.

Comme nous l'avons vu précédemment, une solution possible pour analyser les données de grandes dimensions consiste à faire appel aux méthodes de DMML. Leur utilisation peut être complémentaire à l'utilisation des méthodes de régression dans la recherche d'associations, et d'interactions entre les variables.

L'objectif de cette thèse est l'application des méthodes de DMML à l'analyse des variables explicatives de la symptomatologie dépressive à l'adolescence. En effet, les difficultés soulevées ci-dessus s'appliquent à cette problématique : de très nombreuses variables explicatives de la dépression à l'adolescence ont été identifiées dans la littérature scientifique (paragraphe 1.1.3 page 18). La force de l'association entre ces variables explicatives et la dépression est relativement faible alors que ces variables sont souvent associées les unes avec les autres et qu'il existe de très nombreuses interactions potentielles positives et négatives.

Pour ce faire, parmi les enquêtes récentes s'intéressant aux comportements des adolescents, l'enquête « Processus d'adolescence » (présentée ci-dessous) a été utilisée. Elle avait pour objectif de décrire les adolescents de cette génération et d'analyser leurs comportements en lien avec leur fonctionnement global (données personnelles, environnementales, culturelles…), en leur donnant la parole.

La suite de ce chapitre est dédiée à la description de l'enquête, des données obtenues relatives à la symptomatologie dépressive des adolescents inclus ainsi qu'aux potentielles variables explicatives de cette symptomatologie.

3.2. Matériel

3.2.1. L'enquête : « Processus d'adolescence »

Design et présentation de l'enquête « Processus d'adolescence ».

L'enquête multicentrique transversale en milieu scolaire « Analyse des difficultés des processus d'adolescence aujourd'hui », appelée plus communément « Processus d'adolescence » a eu lieu entre le 12 et le 16 octobre 2013. Cette enquête de type « un jour donné » avait pour objectifs :

- D'analyser les comportements des adolescents dans le contexte culturel actuel en prenant en compte la qualité de leur processus d'adolescence
- De chercher à identifier leurs « profils » à travers de multiples variables et à repérer les indicateurs de difficultés dans le processus adolescent.
- D'analyser l'ensemble des comportements adolescents qui peuvent être considérés comme à risques.

Elle a fait l'objet d'une collaboration entre l'Institut national de la santé et de la recherche médicale (Inserm) U1178 et le pôle universitaire du Centre Hospitalier Spécialisé Fondation Vallée. Elle a été coordonnée en partenariat avec le Ministère de l'Education Nationale et l'Enseignement agricole.

Population étudiée et échantillonnage

Les données ont été collectées auprès d'adolescents scolarisés dans le second cycle, qu'il soit général, technologique, professionnel ou agricole, à partir de la 4ème jusqu'à la terminale (13 à 20 ans, puberté engagée en moyenne ou en grande adolescence). Les adolescents déscolarisés, hospitalisés ou pris en charge et scolarisés dans le médico-social n'ont pas été inclus pour des raisons de faisabilité.

Cette enquête a été menée auprès d'adolescents, issus de territoires géographiques contrastés (urbains, montagnards et ruraux) et éloignés (Ile de France, Ouest et Sud), évoluant dans des contextes sociodémographiques différents.

Six départements différents ont été invités à participer à l'enquête :

-les Hautes Alpes, où toutes les classes de tous les établissements ont été inclues (enquête proposée à tous les adolescents scolarisés et présents le jour de la passation)

- La Vienne, la Charente-Maritime, la Charente et les Deux Sèvres avec seulement les établissements de l'enseignement agricole (lycées et maisons rurales et familiales).

- le Val de Marne dont un échantillon représentatif des adolescents scolarisés dans le cycle secondaire a été tiré au sort (tirage au sort aléatoire à deux niveaux : type d'établissement et classe, réalisé par le service statistique du rectorat de Créteil). La méthode d'échantillonnage est décrite en détail dans le rapport de l'enquête (2).

Au total, 137 établissements scolaires ont été sélectionnés dans les 6 départements afin de participer à l'enquête : parmi eux, 134 ont accepté d'y participer soit 97,8% des établissements initialement prévus.

Ethique

L'ensemble des adolescents des classes sélectionnées, qu'ils soient majeurs ou mineurs ont reçu un formulaire d'information. Ils avaient la possibilité de refuser de participer à l'enquête. Les parents des élèves mineurs, ont également reçu le formulaire d'information et avaient la possibilité de s'opposer à la participation de leur enfant. L'enquête a reçu l'avis favorable du comité consultatif national et de la Commission Nationale de l'Informatique et des libertés (CNIL n°912523). Elle a également reçu l'agrément des associations nationales de parents d'élèves.

Questionnaire

Comme la plupart des enquêtes en milieu scolaire, le recueil des données a été fait *via* un questionnaire anonyme auto administré en classe

Ce questionnaire (Annexe 1) comprenait au total trois cent quarante-huit questions portant sur :

- Les caractéristiques sociodémographiques et la structure familiale
- La scolarité
- La consommation de substances psychoactives
- La perception de l'adolescence
- L'image du corps
- L'alimentation
- Les relations familiales et sociales
- Les loisirs
- La santé

Il s'appuie sur des questionnaires préexistants : l'enquête mise en place par Choquet et Ledoux en 1993 « Adolescents-Enquête Nationale » , et les enquêtes ESPAD (2007) et ESCAPAD (2008), notamment pour les consommations de substances psychoactives (1,19,20) .

Déroulement de l'enquête

Ces auto-questionnaires à choix multiples anonymes ont été proposés aux élèves des classes sélectionnées, durant une unité de cours (50 min) selon une méthodologie standardisée. Les conditions de passation sur l'ensemble des territoires investigués étaient identiques et comparables à celles des grandes enquêtes menées en population adolescente (ESPAD ; ESCAPAD). Les enquêteurs étaient des infirmièr(e)s scolaires et des élèves infirmier(e)s, formés par l'équipe de recherche. Ils disposaient d'un dossier contenant tous les documents nécessaires à la passation ainsi qu'un texte de présentation à lire aux élèves en début de passation afin de standardiser les informations.

Bilan d'enquête

Au total 730 classes ont accepté de participer à l'enquête. Sur l'effectif théorique de 16719 élèves attendus, les motifs de non-participation des élèves correspondaient à un refus des parents (n=46), des élèves (n=26) ou à une absence le jour de l'enquête (n=1370).

Au total, 15277 adolescents ont rendu un questionnaire, soit un taux de participation de 91,4%. Parmi eux, 0,27% des élèves participants (n=42) ont été exclus car leur questionnaire était inexploitable. L'échantillon final était donc constitué de 15235 élèves scolarisés. Les caractéristiques de cet échantillon sont détaillées dans le rapport (2) .

3.2.2. Mesures

3.2.2.1. Définition de la variable d'intérêt clinique : Symptomatologie dépressive

Dans cette étude, le diagnostic de la dépression de l'adolescent ne peut être établi directement. En revanche, il est possible d'évaluer la symptomatologie dépressive et son intensité, grâce à l'« Adolescent Depressive Rating Scale » (ADRS). Il s'agit d'une échelle auto-évaluative spécifiquement crée et validée en Français pour cette tranche d'âge (13-20 ans) par Révah-Lévah et al., (18). Cette échelle est composée de 10 items « vrai/faux » explorant au cours des deux dernières semaines : l'état émotionnel avec irritabilité, l'envahissement par le vécu de la dépression, les perceptions négatives de soi, la présence d'idées noires et les manifestations non émotionnelles (ralentissement physique ; sommeil) (Annexe 2). Le score est déterminé par le nombre de réponses « vraies » avec un total compris entre 0 et 10. Cette échelle montrant une bonne validité factorielle, et une bonne consistance interne, est régulièrement utilisée dans les études épidémiologiques en population adolescente (i.e volet français ESCAPAD, ESPAD).

Elle a également une bonne validité concurrente et discriminante. Elle a été validée en clinique comme outil de repérage des adolescents avec un diagnostic d'épisode dépressif caractérisé (diagnostic DSM-IV effectué par des pédopsychiatres). En 2014, le rapport de la Haute Autorité de Santé, a recommandé aux médecins généralistes d'utiliser le questionnaire ADRS, « test le mieux validé pour aider à la détection d'un épisode dépressif caractérisé ». Un seuil à 6 a été proposé par Révah Lévy et al., (39) pour identifier les adolescents présentant une forte symptomatologie dépressive en accord avec les critères diagnostic d'un épisode dépressif caractérisé selon le DSM-V(13).

Dans la suite de ce travail, les adolescents avec un score supérieur ou égal à 6 seront nommés « adolescents présentant une forte symptomatologie dépressive ».

Par opposition les adolescents ayant un score à l'échelle ADRS strictement inférieur à 6 seront nommés « adolescents présentant une faible symptomatologie dépressive ».

3.2.2.1. Description des variables explicatives disponibles dans l'enquête « Processus d'adolescence »

L'enquête grâce aux 348 questions auxquelles les adolescents ont répondu permet d'explorer le processus d'adolescence dans sa globalité. Parmi celles-ci, nous avons retenu 93 variables incluant des renseignements sociodémographiques (âge, sexe, établissement scolaire) et un ensemble d'items interrogeant des domaines connus dans la littérature pour être des variables explicatives de la dépression de l'adolescent :

1) Caractéristiques sociodémographiques et scolaires : Age, sexe, établissement scolaire, redoublement, sentiment vis-à-vis de l'école.
2) La santé au sens large : sommeil, visites chez le médecin…
3) Consommation de substances psychoactives licites et illicites : expérimentation et consommation régulière intensive au cours des 30 derniers jours de tabac, alcool, ivresse, cannabis, et expérimentation d'autres drogues…
4) La structure familiale : foyer de vie principal (avec au moins un parent), séparation des parents, situation professionnelle des parents (en activité ou retraité versus sans activité), niveau d'études le plus élevé des parents, décès d'un des parents…
5) Alimentation/image du corps : Indice de masse corporelle (calculé à partir du poids et de la taille déclarés par les enquêtés et catégorisés selon les normes de *l'International Obesity Task Force* (91), prenant en compte le genre et l'âge), rapport à l'alimentation…,
6) Relations familiale et sociales : Nombre d'amis dans la réalité et sur internet, discussion avec les parents sur des sujets divers, disputes dans la famille
7) Sexualité : attirance hétérosexuelle ou homosexuelle ou bisexuelle, avoir déjà eu des rapports sexuels, caractéristiques du premier rapport sexuel, avoir déjà eu recours à une interruption volontaire de grossesse.
8) Loisirs : pratique d'un sport (loisir ou compétition, sport à risque), temps de jeux vidéo, faire ou écouter de la musique, lire des livres/Bande Dessinée, dessiner, peindre…
9) Autres : participation à des jeux dangereux

3.2.3. Création de l'échantillon d'analyse

Sur les 15235 adolescents ayant rendu un questionnaire, 917 ont été exclus de mon analyse : les adolescents n'ayant pas renseigné leur genre (n=4) ; les adolescents dont les questionnaires détenaient plus de 33% de données manquantes sur l'intégralité de leurs réponses (n=171) ; les adolescents âgés de moins de 13 ans (n=75) ou de plus de 20 ans (n=38) ou n'ayant pas renseigné leur âge (n=459) ; ceux dont le score à l'ADRS n'a pas pu être calculé (n=170).

Les analyses transversales de la relation entre la symptomatologie dépressive (variable d'intérêt clinique) et les différentes variables explicatives ont donc été effectuées sur 14318 adolescents scolarisés en collège, lycée professionnel ou agricole ou lycée général et technologique, pour qui, la symptomatologie dépressive a pu être définie. L'ensemble des analyses a été réalisé séparément chez les garçons (n=6805) et les filles (n=7513), compte-tenu des différences bien établies de prévalence de la dépression et de comportements entre les sexes (92,93) .

3.2.4. Population analysée

La population analysée compte 14318 adolescents au total. Le Tableau 5, résume leurs caractéristiques sociodémographiques.

Notre échantillon est composé de 52% de filles (N=7513), âgées en moyenne de 15,5 ans ± 1,6 et 48% de garçons âgés en moyenne de 15,4 ans ± 1,6 (N=6805). Ces adolescents étaient majoritairement scolarisés en lycée (65,08%).

Concernant la symptomatologie dépressive, 12,13% de l'échantillon total présentait une forte symptomatologie dépressive (score ADRS ≥6) : 16,8% (N=1263) des filles et 6,9% des garçons (N=474). Dans l'enquête ESPAD 2007, la prévalence d'une forte symptomatologie dépressive chez les filles était de 11,4% contre 5,4% chez les garçons (21). Dans l'enquête ESCAPAD 2008, la prévalence d'une forte symptomatologie dépressive chez les filles était de 10,4% contre 4,5% chez les garçons (22).

Tableau 5: Description de l'échantillon total

Caractéristiques sociodémographiques	Total N=14318	Filles N=7513	Garçons N=6805
Region:			
Rurale	6654 (46,47%)	3409 (45,37%)	3245 (47,69%)
Montagnarde	3820 (26,68%)	2011 (26,77%)	1809 (26,58%)
Urbaine	3844 (26,85%)	2093 (27,86%)	1751 (25,73%)
Age compris entre:			
[13-15]	4346 (30,35%)	2207 (29,38%)	2139 (31,43%)
[15-18[	8561 (59,79%)	4481 (59,64%)	4080 (59,96%)
[18-20]	1411 (9,85%)	825 (10,98%)	586 (8,61%)
AGE (moy±E-T)	15,46 ±1,61	15,53 ±1,64	15,38 ±1,57
Type d'établissement scolaire:			
Collège	4944 (34,92%)	2442 (32,81%)	2502 (37,25%)
Lycée général et technologique	5144 (36,33%)	2896 (38,91%)	2248 (33,47%)
Lycée professionnel et agricole	4072 (28,76%)	2105 (28,28%)	1967 (29,28%)
Redoublement	4459 (31,18%)	2261 (30,13%)	2198 (32,35%)
Vivre avec au moins un parent	12590 (89,56%)	6616 (89,39%)	5974 (89,75%)
Décès d'au moins un parent	670 (4,81%)	363 (4,93%)	307 (4,68%)
Niveau d'étude le plus élevé du père (Niveau Baccalauréat)	5056 (45,86%)	2471 (43,94%)	2585 (47,85%)
Niveau d'études le plus élevé de la mère (Niveau Baccalauréat)	6623 (57,27%)	3254 (53,79%)	3369 (61,10%)
Symptomatologie dépressive			
Faible (score <6)	12581 (87,87%)	6250 (83,19%)	6331 (93,03%)
Forte (score>=6)	1737 (12,13%)	1263 (16,81%)	474 (6,97%)

3.2.5. Associations entre la symptomatologie dépressive et variables explicatives

Une analyse bivariée a été réalisée afin d'évaluer l'association brute entre la symptomatologie dépressive et les différentes variables explicatives. Ces analyses ont été réalisées à l'aide de tests de $\chi 2$ dont les résultats sont présentés dans l'Annexe 3.

Bien qu'évoluant dans des territoires géographiques contrastés (urbain, montagnard et rural) la prévalence de la symptomatologie dépressive n'était pas significativement différente entre les régions. De même, il n'y avait pas de différence statistiquement significative de la prévalence d'adolescents avec une forte symptomatologie dépressive selon les groupes d'âge, l'établissement (collège/lycée) ou la filière. En revanche, quasiment toutes les autres variables explicatives s'avéraient statistiquement associées à une forte symptomatologie dépressive chez les filles et/ou les garçons. Seules six variables se sont avérées sans association statistiquement significative avec la symptomatologie dépressive à la fois chez filles et chez les garçons : *« Etes-vous attentif à votre physique ? », « la scolarité est selon vous la seule chose qui compte pour vos parents »*, *« Fréquence des loisirs préférés : jouer à des jeux de*

sociétés », « pratique d'un sport à risque », « avoir un ordinateur personnel » et « jouer à des jeux sur un ordinateur ».

Comme attendu, les adolescents avec une forte symptomatologie dépressive ont plus que les autres, tendance à avoir redoublé et à ne pas aimer pas l'école. Ils ont tendance à avoir expérimenté des substances psychoactives, consommé régulièrement du tabac, de l'alcool ou du cannabis (lors du mois précédent l'enquête) et à avoir ressenti une intensité d'ivresse plus élevée la dernière fois qu'ils ont bu. Ils sont plus susceptibles d'être décalés (s'endormir tard, se réveiller tard) ; d'avoir été en surpoids, de ne pas faire attention à leur alimentation et de ne pas aller régulièrement voir un médecin.

Ils ont une situation familiale qui diffère des adolescents avec une faible symptomatologie dépressive : ils sont plus nombreux à avoir des parents divorcés, un père ou une mère n'ayant pas le baccalauréat ou sans activité professionnelle. Ils ont plus tendance à rapporter des disputes dans leur famille, et ne pas parler facilement avec leurs parents de l'école, de leur santé, de leurs problèmes. Ils ont également plus tendance à participer à des jeux dangereux, à pratiquer un sport à risque et à avoir eu un premier rapport sexuel sans protection.

La relation aux pairs est également statistiquement associée à la symptomatologie dépressive : les adolescents avec une forte symptomatologie dépressive déclarent plus avoir une attirance sexuelle homosexuelle et bisexuelle, avoir plus d'amis sur internet, et avoir peu d'amis dans la réalité.

Leur type de loisir préféré est également significativement différent entre les deux groupes d'adolescents, avec, par exemple, plus de sujets dans le groupe à une forte symptomatologie dépressive, qui préfèrent utiliser un ordinateur pour l'internet et jouer de manière intensive (plus de 3h par jour) aux jeux vidéo durant la semaine/le week-end ou les vacances, mais dans le groupe des sujets à faible symptomatologie dépressive ; il y a plus de jeunes qui mentionnent comme loisir préféré faire du sport en loisir ou en club ou être avec des amis de la réalité (dessin autres …..).

Les données issues de l'enquête présentée ci-dessus, seront utilisées dans les deux prochains objectifs de thèse. Dans un premier temps, il s'agira d'évaluer l'intérêt des méthodes d'agrégation d'arbres par rapport à une méthode de régression pénalisée afin d'analyser l'association entre les variables explicatives de la symptomatologie dépressive l'adolescent ; dans un second temps, l'objectif sera d'appliquer une méthode de partitionnement supervisée par la variable d'intérêt clinique, afin d'identifier et de caractériser des profils différents d'adolescents à risque de présenter une forte symptomatologie dépressive.

4. Intérêt des méthodes DMML de classification à l'analyse de l'association entre la symptomatologie dépressive à l'adolescence et ses variables explicatives.

Statistiquement il s'agit de modéliser une variable d'intérêt clinique qui suit une distribution binomiale à partir de variables explicatives catégorielles et quantitatives. Comme nous l'avons exposé précédemment, de nombreuses variables explicatives sont à tester ; ces variables sont globalement associées les unes aux autres et la prise en compte de multiples interactions, d'ordre deux ou plus, semble prometteuse. Classiquement en épidémiologie, un modèle de régression logistique serait utilisé pour ce modèle. Néanmoins, avec 93 variables, cela implique d'envisager un espace de 4278 paires d'interactions possibles, et 129766 interactions d'ordre 3. Si l'échelle reste bien inférieure à celles des « Big Data » de type omique, le nombre de tests n'en est pas moins bien supérieur au nombre de sujets, la sélection des variables (et des interactions) et leurs estimations délicates et cela rend particulièrement attractive l'appel à des méthodes alternatives.

Pour ce faire, deux méthodes d'agrégation d'arbres : la Forêt aléatoire (RF) et Descente de gradient stochastique (SGD), et une méthode de régression logistique régularisée par pénalisation par LASSO, ont été comparées. Le modèle de régression LASSO, de plus en plus populaire en épidémiologie se situe à la frontière entre statistique classique et le « Machine Learning ». Il a été utilisé, comme méthode de référence en comparaison avec les méthodes d'agrégation d'arbres ; cette méthode adaptée aux problèmes de grande dimension a pour avantage d'effectuer une sélection de variables et de fournir des coefficients directement interprétables. Au contraire, les méthodes d'agrégation d'arbres sont clairement identifiées comme faisant partie des méthodes de « Machine Learning ». Elles ne fournissent pas directement de coefficients interprétables mais ont pour avantage de prendre en compte implicitement les interactions sans avoir à les spécifier préalablement (94). En effet, la structure hiérarchique d'un arbre implique que la réponse à une variable explicative dépend des valeurs situées précédemment dans l'arbre, de sorte que les potentielles interactions entre les prédicteurs soient approximées (94–96). De plus, les méthodes d'agrégation d'arbres dépendent moins des propriétés de l'ensemble de données, en particulier des corrélations entre les prédicteurs que les régressions LASSO (voir méthodes LASSO paragraphe 4.1.1 page 69)

L'hypothèse de départ était que les méthodes d'agrégation d'arbres, permettraient une prédiction de la présence d'une symptomatologie dépressive plus efficace que la méthode LASSO. L'objectif secondaire de cette analyse était de comparer les variables importantes dans les différents modèles de façon à obtenir des informations complémentaires sur les patterns de variables explicatives identifiant les adolescents avec une forte symptomatologie dépressive. J'ai donc développé trois modèles indépendants (régression pénalisée LASSO, forêt aléatoire : RF et arbre de régression boostée

de type Descente de gradient stochastique : SGD) sur un même échantillon d'apprentissage issu de l'enquête « Processus d'adolescence ». La qualité des trois modèles a été comparée sur un autre échantillon de l'enquête (échantillon de validation) et les variables majeures des modèles d'agrégation d'arbre ont été analysées en regard des variables majeures de la régression LASSO.

4.1. Matériel et Méthode

4.1.1. Régression Lasso

La régression pénalisée LASSO est une extension des modèles de régression linéaires généralisées. Le LASSO est adapté aux problèmes de grande dimension (n << p) et s'appuie sur des algorithmes peu coûteux en temps de calcul et stockage. Ce type de régression fait partie des méthodes dites de « shrinkage » dont l'objectif est de conserver les variables explicatives ayant une force d'association avec la variable d'intérêt les plus importantes et d'ignorer celles avec une force d'association minimale. Cette technique permet d'étudier l'ensemble des variables simultanément et convient aux modèles présentant des niveaux élevés de multi-colinéarité (97) . Elle minimise également le risque de sur-ajustement.

La méthode LASSO estime les coefficients de régression par maximisation de la log-vraisemblance du modèle, tout en forçant la somme des valeurs absolues de ces coefficients de régression à être inférieure à une valeur fixe. Le choix du paramètre λ de régularisation est un élément fondamental de l'utilisation du LASSO qui contrôle la force du retrait des variables. En effet, pour un λ nul, aucune pénalité n'est appliquée; pour une valeur de λ non nul, le LASSO réduit certains des coefficients estimés à zéro (98) .

En pratique, la régression LASSO effectue une sélection automatique des variables puisque les variables associées avec un coefficient égal à 0, sont de facto exclus du modèle prédictif. Néanmoins, la priorisation dans le processus de sélection peut dépendre fortement des propriétés de l'ensemble de données avec une tendance à ne sélectionner qu'une seule variable explicative parmi celles qui sont fortement corrélées. En outre, il n'est pas certain que les variables sélectionnées soient systématiquement celles ayant la plus forte association avec la variable à prédire (99).

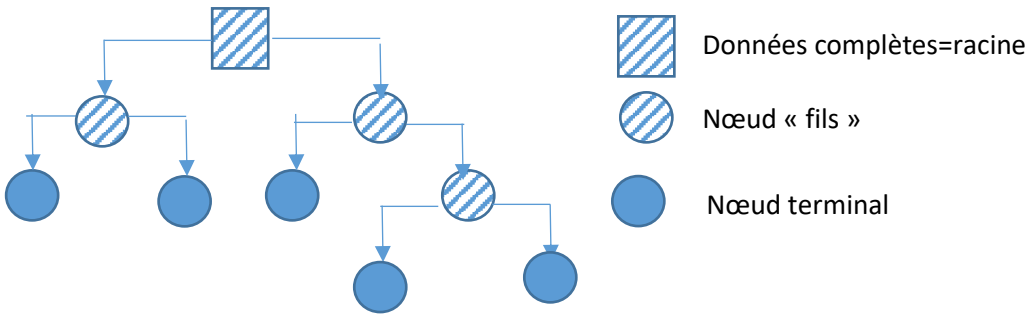
4.1.2. Méthodes d'agrégation d'arbres

4.1.2.1. Principe général de construction d'un arbre

Les méthodes de forêt aléatoire (RF) et de descente de gradient stochastique (SGD) reposent sur l'agrégation de modèles d'arbre de décision de type CART.

Un arbre donné est construit de la manière suivante : en partant des données complètes (i.e la racine), un nœud est défini par le choix conjoint d'une variable parmi toutes les variables explicatives et d'une division (définie par une valeur seuil si la variable est quantitative ou un partage en deux groupes des modalités si la variable est qualitative). Les observations sont ainsi partitionnées en deux groupes composants chacun un nœud « fils ». La variable et la division définissant la coupure sont choisies de façon à construire deux nouveaux nœuds les plus homogènes possible au sens de la variable à expliquer. Le partitionnement est répété jusqu'à ce que les derniers nœuds ne contiennent que des sujets ayant la même valeur en ce qui concerne la variable à expliquer ou jusqu'à ce qu'ils ne contiennent qu'un seul sujet (Figure 9). Un arbre final est ainsi obtenu et un élagage peut être appliqué pour éviter le sur-apprentissage. Les nœuds terminaux également appelés feuilles sont utilisés comme prédictions (100).

Figure 9: Construction d'un arbre de classification



La structure hiérarchique d'un arbre signifie que la réponse à une variable dépend des valeurs des variables situées plus haut dans l'arbre, de sorte que les interactions éventuelles entre les variables prédicteurs sont automatiquement modélisées.

4.1.2.2. Forêts aléatoires (RF)

Une forêt aléatoire est par définition construite par une agrégation de plusieurs centaines, voire plus, d'arbres construits indépendamment sur des sous-échantillons différents de l'échantillon d'apprentissage. Concrètement l'élaboration de chaque arbre commence par la constitution d'un sous-échantillon par tirage au sort avec remplacement sur les données d'apprentissage puis l'algorithme fait croitre l'arbre sur cet échantillon en répétant récursivement les différentes étapes à chaque nœud de l'arbre (100,101) :

1) Tirage au sort de « m variables » (également appelé mtry) parmi toutes celles possibles.
2) Parmi ces « m variables » : choix de celles permettant de minimiser le critère d'impureté de Gini (un indice permettant d'évaluer l'hétérogénéité des nœuds)
3) Division du nœud en deux nœuds fils.

La procédure de sélection aléatoire des variables à chaque nœud, assure une faible corrélation entre les arbres ce qui permet de prévenir les problèmes de sur-apprentissage. Une fois la forêt aléatoire développée, la prédiction globale pour un nouvel individu correspond à la classe majoritaire prédite par l'ensemble des arbres construits.

Dans une forêt aléatoire, tous les arbres sont différents, c'est l'agrégation d'arbres taillés via des variables distinctes qui fournit les capacités prédictives. De fait, l'algorithme n'aboutit pas un arbre moyen lisible qui permettrait d'extrapoler de la connaissance sur les prédicteurs et leurs éventuelles interactions. Cependant, il est possible d'estimer l'importance des variables dans la classification des individus, ce qui permet d'identifier les prédicteurs les plus importants.

4.1.2.3. Descente de gradient stochastique (Stochastic Gradient Boosting Model SGD)

Alors que le modèle RF est un ensemble d'arbres de décision construits indépendamment les uns des autres, le modèle SGD est un modèle additif qui construit un ensemble d'arbres de manière récursive. Le « boosting » est une méthode permettant d'améliorer la précision des modèles, basée sur l'idée qu'il est plus facile de trouver et de faire la moyenne de nombreuses règles approximatives que de trouver une seule règle de prédiction très précise (102). Dans le cas des arbres boostés, le modèle combine séquentiellement des arbres de classification « simples » (constitués d'un nombre limité de nœuds terminaux) pour produire une règle de classification précise. En d'autres termes, l'objectif de formation de chaque nouvel arbre est de minimiser les écarts entre les résultats observés et ceux prévus par tous les arbres précédents (103). L'objectif de formation de chaque nouvel arbre est de minimiser les écarts entre les résultats observés et ceux prévus par tous les arbres précédents (103).

Schématiquement, à la première étape de l'algorithme un arbre de classification est construit et des résidus de prédiction de l'arbre sont calculés. Le second arbre est construit sur les résidus du premier arbre. Les résidus du modèle contenant les deux arbres sont calculés, un nouvel arbre est construit, et ainsi de suite. Pour réduire le sur-apprentissage, chaque arbre de régression est fondé à partir d'un sous-échantillon différent du jeu de données d'apprentissage construit aléatoirement (sans remise) puis l'erreur du modèle est estimée en prédisant la variable réponse chez les sujets non inclus dans cet échantillon. Les itérations sont poursuivies jusqu'à ce que l'amélioration des performances prédictives soit considérée comme marginale ou lorsqu'un un nombre maximum d'arbres a été atteint. Une fois l'ensemble d'arbre construit, un vote pondéré sur les décisions des différents arbres est utilisé lors de la prédiction d'un nouvel individu.

Comme pour les modèles RF, le modèle inclut des centaines d'arbres différents, ce qui, de fait, induit un manque d'interprétabilité du modèle. Toutefois, il est possible d'estimer numériquement l'importance relative de chacune des variables incluses pour prédire la variable réponse.

4.2. Plan d'analyse

4.2.1. Echantillons d'apprentissage et de validation externe

Les trois modèles (LASSO, RF et SGD) ont été créés sur le même échantillon d'apprentissage, obtenu par tirage aléatoire sans remise à partir de l'ensemble de données globales de façon à contenir 75% des sujets. Réciproquement, un échantillon d'évaluation (pour la validation externe) a été réalisé, contenant le quart restant de la population. Cet échantillon de validation, a servi à comparer les performances des modèles LASSO, SGD et RF. La création des échantillons d'apprentissage et d'évaluation a été stratifiée sur la prévalence d'adolescents présentant une forte symptomatologie dépressive (Tableau 6).

Tableau 6: Echantillonnage

Sexe	N validation	N évaluation	Total
Filles	5636 (75%)	1877 (25%)	7513
Garçons	5105 (75%)	1700 (25%)	6805

4.2.2. Construction des modèles via le package « CARET »

Lors de la construction d'un modèle, quel que soit l'algorithme utilisé (SGD, RF ou LASSO), il est nécessaire de régler des hyperparamètres lors du processus d'apprentissage. Les hyperparamètres sont des paramètres structurels qui définissent l'architecture du modèle. Ainsi, dans un arbre de régression boosté de type SGD, les hyperparamètres incluent la profondeur maximale de chaque arbre, le nombre d'arbres, le taux d'apprentissage et le nombre minimal d'observations dans les nœuds terminaux des arbres.

L'approche pragmatique pour optimiser les hyperparamètres, consiste à essayer les algorithmes avec différentes combinaisons de paramètres, comparer leurs performances, et sélectionner l'architecture qui produit les meilleurs résultats pour l'analyse finale. Comme les mesures de performance peuvent être sensibles à la partition sur laquelle elles sont mesurées, le choix des hyperparamètres est pris sur une moyenne de performances de plusieurs échantillonnages du jeu de données afin d'éliminer les biais liés à la structure d'un ensemble d'évaluation. Pour ce faire, la librairie Caret (**C**lassification **A**nd **RE**gression **T**raining) implémentée sur le logiciel R a été utilisée (104). Cette librairie a de nombreuses fonctions qui permettent d'optimiser le processus de construction et d'évaluation du modèle : préparation des données, réglages par ré-échantillonnage des hyperparamètres du modèle, estimation des performances du modèle optimal sur un ensemble de données externes, comparaison du poids des variables incluent dans un modèle.

Le protocole d'optimisation des hyperparamètres pour chaque algorithme suit les étapes suivantes et sont résumés en Figure 10 :

1) A partir des valeurs des hyperparamètres proposées par les utilisateurs, une grille est générée définissant toutes les combinaisons possibles qui seront testées. (Les hyper paramètres testés sont présentés dans le Tableau 8 pages 79).
2) Pour tester chaque combinaison nous avons utilisé une approche par validation répétée et croisée à k blocs, « repeated k-fold cross-validation ».
    a. L'échantillon d'apprentissage fourni à Caret est divisé aléatoirement en cinq blocs de taille égale. Chacun des blocs servira tour à tour d'ensemble de test, pendant que la combinaison des quatre autres blocs constitue l'ensemble d'entrainement. La procédure ci-dessus est répétée dix fois avec dix tirages au sort différents, créant ainsi 50 couples ensembles d'entrainement / ensemble de test à partir de l'échantillon originel. Dans notre population, la distribution de la variable à expliquer est très déséquilibrée (en particulier chez les garçons), une telle disparité dans les fréquences

des classes observées peut avoir un impact négatif important sur l'ajustement des modèles de machine Learning. Il est donc conseillé d'échantillonner les ensembles d'apprentissage de manière à atténuer le déséquilibre. Pour ce faire, nous avons artificiellement multiplié les individus avec une forte symptomatologie dépressive dans chaque ensemble d'apprentissage *via* un tirage au sort (avec remplacement) de ces sujets minoritaires de façon à ce que la classe minoritaire ait la même taille que la classe majoritaire.

b. 50 modèles sont construits (un sur chacun des échantillons d'entrainement) et ces modèles sont testés sur les échantillons de test correspondant. Le critère de performance utilisé ici est le Kappa couramment utilisé en Machine Learning pour évaluer des modèles de classifications et particulièrement indiqué lorsque les classes sont fortement déséquilibrées. La performance globale de la combinaison d'hyperparamètres est évaluée en agrégeant les 50 Kappa obtenus sur les 50 échantillons de test.

3) Les Kappa globaux de chaque combinaison sont comparés pour déterminer quelle combinaison des paramètres est optimale.
4) Les valeurs optimales sont attribuées, le modèle final est réajusté en utilisant le set d'apprentissage.

4.2.3. Comparaison de la performance des 3 modèles (LASSO, SGD, RF)

Une fois les trois modèles optimisés sur l'échantillon d'apprentissage, il s'agit de comparer leur performance. Les métriques classiques en statistique de comparaison de modèles intégrant la complexité du modèle, sont mal adaptées à une philosophie « Big data » visant à incorporer un maximum d'informations. L'approche DMML suppose que la relation entre la variable d'intérêt réponse et les prédicteurs est complexe et inconnue, sans postuler de modèles (e.g sans postuler de distribution de la variable réponse) et tente d'apprendre cette relation en observant les données. Cela met l'accent sur la capacité d'un modèle à bien prédire, et se concentre sur ce qui est prédit et comment le succès de la prédiction doit être mesuré (94). En conséquence, lors des approches DMML, la comparaison des modèles de classification est traditionnellement effectuée par mesure des performances de prédiction du modèle sur des observations non utilisées lors de l'apprentissage. Le taux de succès, le Kappa de Cohen, la sensibilité ou la spécificité sont ainsi des métriques couramment utilisées pour évaluer la qualité du modèle, même lorsque l'objectif des modélisations n'est pas, comme ici, la construction d'un algorithme de prédiction automatique :

- le coefficient Kappa (105) , très utilisé en « Machine Learning », évalue la concordance entre la classe réelle et la prédiction afin de définir le meilleur algorithme. Il dépend à la fois de la

concordance observée (Po, la proportion d'agrément observé) et de la concordance attendue par hasard concordance calculée Ph (proportion d'agrément attendu par hasard attendue sous l'hypothèse d'indépendance).

$$K = \frac{\mathrm{Po} - \mathrm{Ph}}{(1 - Ph)} = \frac{2(bc - \mathrm{ad})}{\big((c + d) \times (a + c)\big) + ((a + b) \times (b + d))}$$

Sa valeur varie entre 0 (désaccord total) et 1 (accord parfait).

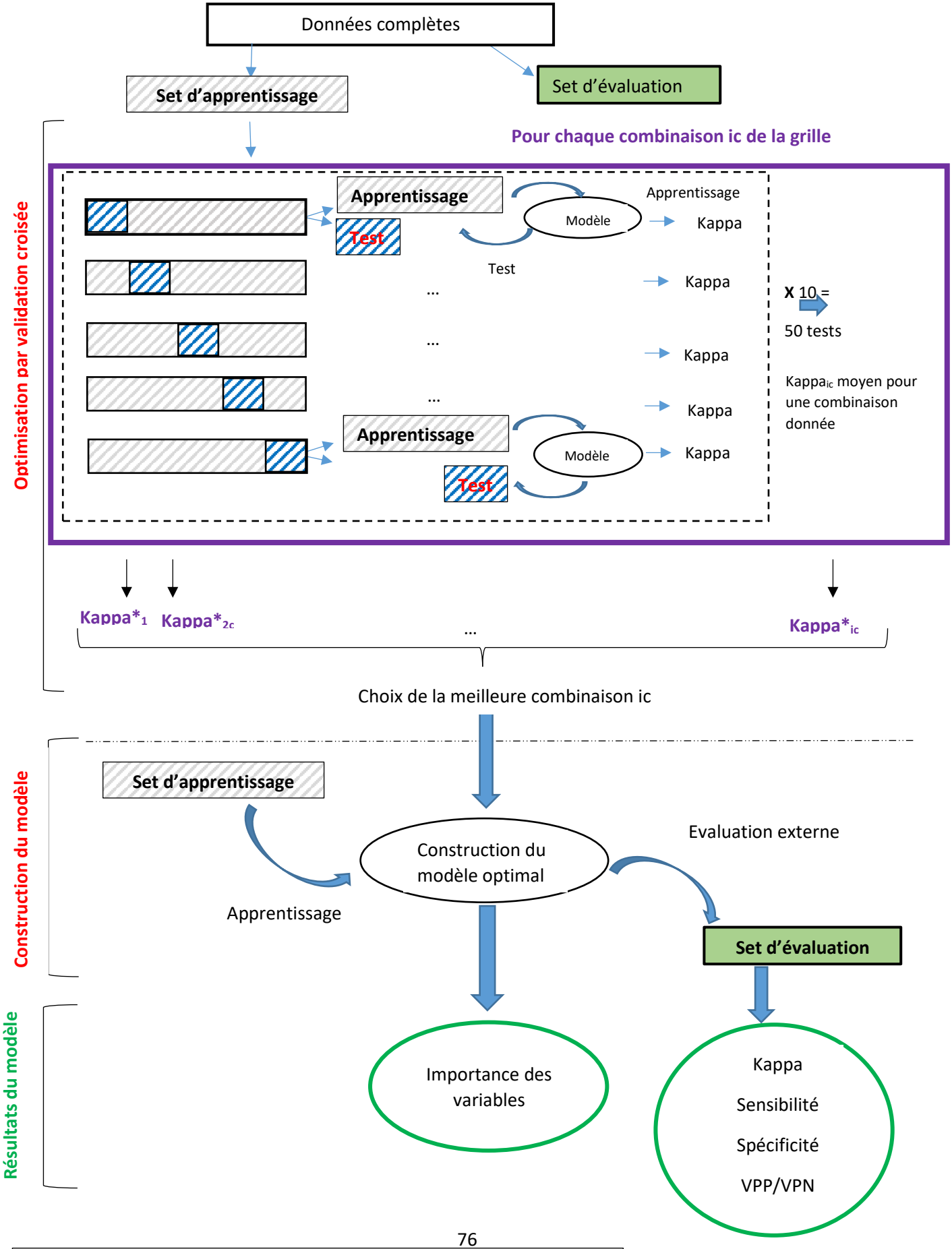
- la sensibilité est le taux de positifs parmi tous les sujets à forte symptomatologie dépressive $Se = \frac{\mathrm{d}}{\mathrm{b+d}}$
- la spécificité est le taux de négatifs parmi tous les sujets à faible symptomatologie dépressive $\mathrm{Sp} = \frac{a}{\mathrm{a+c}}$
- la valeur prédictive positive est la probabilité qu'un sujet à forte symptomatologie dépressive soit prédit positif $VPP = \frac{\mathrm{d}}{\mathrm{c+d}}$
- la valeur prédictive négative est la probabilité qu'un sujet à faible symptomatologie dépressive soit prédit négatif $VPN = \frac{\mathrm{a}}{a+b}$

Tableau 7: Table de contingence Prédiction/Données réelles

	Faible symptomatologie dépressive	**Forte symptomatologie dépressive**	**Total**
Prédiction négative par le modèle	a	b	a+b
Prédiction positive par le modèle	c	d	c+d
Total	a+c	b+d	N

L'utilisation d'une métrique telle que le taux de succès, lorsque, comme ici, la grande majorité des observations appartient à la même catégorie (i.e les adolescents à faible symptomatologie dépressive), privilégiera un modèle peu « intelligent » prédisant quasi systématiquement la classe dominante. Les critères de jugement de performance choisis ont donc été le Kappa de Cohen qui intègre la concordance attendue par hasard, ainsi que la VPP et la sensibilité qui se focalisent sur le nombre de vrais positifs, VPN et spécificité seront donnés à titre indicatifs mais non utilisées pour classer les modèles (Tableau 7).

Figure 10: Processus de création d'un modèle



Kappa*ic= kappa moyen pour une combinaison donnée

4.2.4. Analyse des variables importantes

La régression LASSO a pour avantage de sélectionner un sous-ensemble restreint de variables (dépendant du paramètre λ et de fournir des coefficients qui peuvent être utilisés comme des métriques d'importance des variables correspondantes dans le modèle (104). Par contre, la régression LASSO ne calcule pas de p-valeur. En conséquence, un coefficient non nul ne signifie pas forcément que la variable retenue serait statistiquement associée à la variable réponse au sens classique du terme ; (e.g dans un modèle de régression logistique au seuil de 5% (99,106)).

Un avantage de l'utilisation isolée des arbres de décision est leur simplicité d'interprétation, mais cet atout est perdu dans les modèles d'agrégation, qui contiennent des centaines voire des milliers d'arbres. De plus, les modèles d'agrégation ne fournissent pas de coefficients de régression, ni de p-valeur. Pour pallier ce manque d'interprétabilité, un certain nombre de méthodes ont été développées afin de déterminer quelles variables sont importantes dans un tel modèle et comment elles influencent les prédictions. Une méthode agnostique a été choisie – c'est-à-dire ne reposant pas sur la structure du modèle. Elle peut donc être appliquée à tout modèle (y compris la régression LASSO) et permet de comparer l'importance d'une variable explicative entre des modèles de structures différentes. Le principe introduit par Breiman en 2001 et généralisé par Fisher et al., est de mesurer combien la performance d'un modèle chute si l'effet d'une variable explicative, ou d'un groupe de variables, est supprimé (101,107,108). Pour supprimer l'effet, les valeurs de la variable sont permutées (puisqu'en permutant les valeurs, toute relation entre la variable explicative et la variable à expliquer est détruite). Si une variable explicative est importante, on s'attend à ce qu'après permutation des valeurs de la variable, la performance du modèle se détériore. Plus la variable est importante, plus la variation de performance est importante. La mesure de l'importance d'une variable explicative prend automatiquement en compte toutes les interactions impliquant la variable en question (109). Cela signifie que l'estimateur de l'importance d'une variable inclut à la fois l'effet principal de la variable sur les performances du modèle et les effets d'interaction entre cette variable et d'autres variables explicatives.

Sur ce principe, l'importance des variables explicatives des trois modèles a été estimée par la différence de la valeur du coefficient kappa obtenue sur les données d'origine et celle obtenue après avoir permuté les valeurs de la variable explicative. $\Delta_{kappai}=Kappa_{originei}-Kappa_{permutéi}$

Avec $Kappa_{permutée}$*=Kappa obtenu par le modèle sur les données permutées pour la co-variable i, et* $Kappa_{originei}$*= Kappa obtenu par le modèle sur les* données d'origine*.*

L'utilisation d'une permutation des données implique un caractère aléatoire et des divergences de structure selon la permutation effectuée. Ainsi, les résultats peuvent dépendre de la configuration obtenue des valeurs permutées. La procédure a donc été répétée 50 fois et les Δ_{kappa} moyennés sur l'ensemble des permutations ont été obtenus. De cette façon, l'incertitude associée aux valeurs calculées quant à l'importance des variables peut être évaluée via un intervalle de confiance à 95% (109). Les variables dont l'intervalle de confiance à 95% du Δ_{kappa} ne comprenait pas 0 ont été considérées comme des variables explicatives « significatives » du modèle.

En pratique, l'approche a été implémentée en utilisant le package « iml » (110) avec le logiciel R et procède comme suit :

1) Calcul du Kappa global du modèle sur les données d'apprentissage d'origine ($Kappa_0$)
2) Pour la variable i
    a) Permutation aléatoire des valeurs de cette variable dans l'échantillon d'apprentissage
    b) Application du modèle sur la base de données avec la variable i « permutée » et estimation des performances prédictives du modèle ($Kappa_i$)
    c) Estimation de l'importance de la variable par le calcul de la différence entre Kappa d'origine et $Kappa_i$ permuté (Δ_{kappa})
    d) Répétition des étapes a,b,c, cinquante fois avec des permutations différentes, afin d'obtenir une distribution de Δ_{kappa}

Cette approche de l'estimation de l'importance des variables a permis de :

i. Déterminer les 10 prédicteurs les plus importants de la symptomatologie dépressive (avec les valeurs de Δ_{Kappa} les plus élevés) selon chaque modèle.
ii. Comparer l'importance des variables explicatives entre des modèles de structures différentes.

4.3. Résultats

4.3.1. Optimisation des modèles

Le Tableau 8 présente les différents hyperparamètres testés afin d'obtenir les modèles optimaux. Les valeurs retenues et appliquées à chaque sous population y sont également présentées.

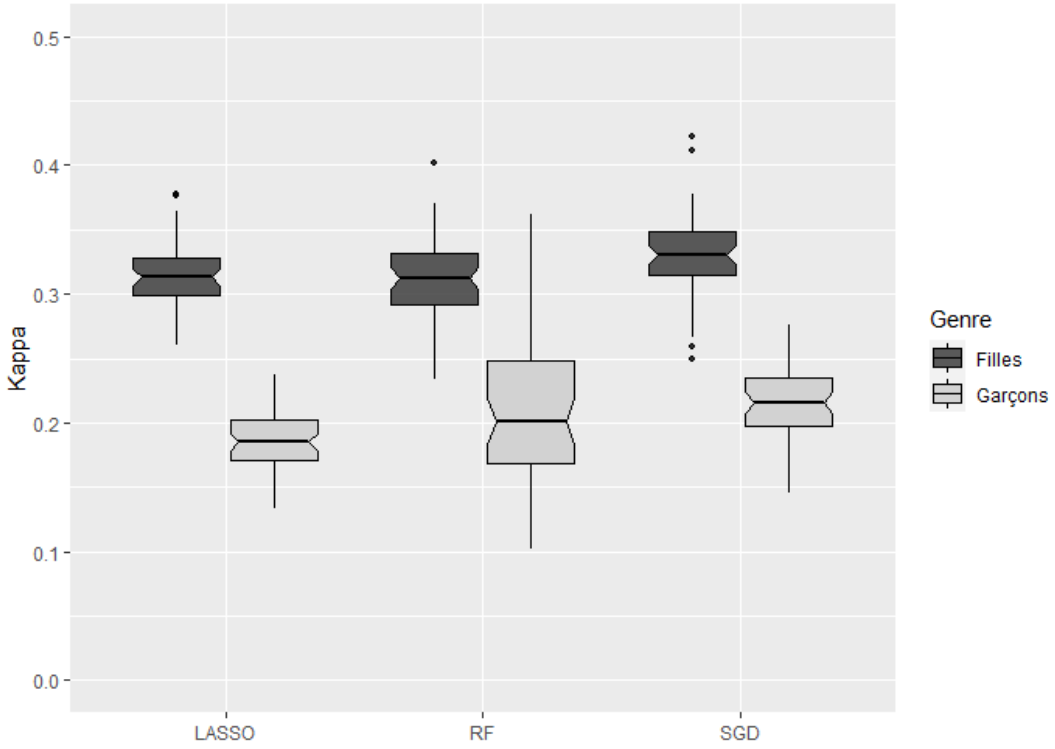
Tableau 8: Description des hyperparamètres testés et valeur retenue par genre

	Hyper-paramètre testés	**Valeur retenue (Filles)**	**Valeur retenue (Garçons)**
Lasso	Lambda de 0,0001 à 1	0,01014925	0,01517387
Foret aléatoire (RF)	mtry de 1 à 15	1	1
	Nombre d'arbres	500	500
Descente de gradient stochastique (SGD)	Nombre d'arbres de 250 à 1000	250	1000
	Nombre de nœuds maximum par arbre de 2 à 10	10	8
	Paramètre de rétrécissement 0,01 à 0,4	0,05	0,01
	Nombre minimum d'observations	10	10

Légende : mtry=nombre de variables tirées au sort

La Figure 11 montre la distribution des 50 Kappas obtenus avec la combinaison optimale d'hyper paramètres de chaque modèle pendant la phase de validation croisée de l'étape de construction. Comme le montre la longueur des boites à moustache, les valeurs de Kappa fluctuaient largement selon le jeu de donnée de validation croisée.

Figure 11: Distribution des Kappa obtenus avec les combinaisons optimales d'hyperparamètres sur les 50 sets d'évaluation



Les moyennes (E-T) de ces Kappas étaient de 0,3161 (0,0255) pour les modèles de régression LASSO, de 0,3123 (0,0318) pour les RF et de 0,3319 (0,0332) pour les SGD sur les données des filles de l'échantillon. Sur les données des garçons les moyennes (E-T) de ces Kappas étaient de 0,1856 (0,0249) pour les modèles de régression LASSO, de 0,2090 (0,0527) pour les RF et de 0,2149 (0,0287) pour les SGD.

4.3.2. Evaluations des performances des modèles

La comparaison des valeurs de Kappa, sensibilité et VPP sur l'échantillon de validation a été utilisée comme technique empirique d'évaluation de la qualité des trois modèles en mesurant leur capacité à se généraliser à une base de données indépendante de la base utilisée pour leur apprentissage.

Le Tableau 9 résume les indicateurs de performances obtenus pour les trois modèles LASSO, SGD et RF. Les modèles filles ont obtenu des Kappa sur le set de validation externe d'environ 0,30 quel que soit l'algorithme utilisé ; les modèles LASSO et SGD étant de 0,31 et celui du RF de 0,29. Chez les garçons, les performances en termes de Kappa étaient légèrement inférieures ; le coefficient Kappa le plus élevé obtenu par le modèle SGD atteignait 0,24 alors que les modèles RF et LASSO ont obtenu

respectivement des valeurs de Kappa de 0,20 et 0,19. Ces valeurs sont tout à fait cohérentes avec celles obtenus en validation croisée lors du processus de construction du modèle (Figure 11).

Similairement les VPP obtenus fluctuaient relativement peu selon les types d'algorithmes mais étaient supérieures chez les filles. Les modèles LASSO ont montré une VPP de 35% chez les filles et de 18% chez les garçons, la VPP était légèrement améliorée par les modèles d'agrégation d'arbre atteignant avec les modèles RF 38% chez les filles et 24% chez les garçons. Par contre, les valeurs de sensibilité obtenues différaient selon les types d'algorithmes. Quel que soit le genre, les modèles LASSO, ont détecté environ 70% des adolescents avec une forte symptomatologie dépressive, ($Se_{LASSOfilles}$=0,69 ; $Se_{LASSOgarçons}$=0,70), soit une sensibilité supérieure à celle des modèles SGD ($Se_{LASSOfilles}$=0,62 ; $Se_{LASSOgarçons}$=0,59) ; les modèles RF ont quant à eux obtenu des valeurs de sensibilité très faibles ne détectant que 46% des filles et 27% des garçons avec une forte symptomatologie dépressive.

Les modèles LASSO obtenaient donc la meilleure sensibilité et les modèles RF la meilleure VPP, néanmoins l'avantage en termes de sensibilité (Delta-filles= 0,23 ; Delta-garçons=0,43) du modèle LASSO sur le modèle RF est largement supérieur au préjudice en termes de VPP (Delta-filles= -0,03 ; Delta-garçons= -0,06). On peut noter que, comme attendu du fait du déséquilibre (de la variable prédite) en faveur des sujets sans symptomatologie dépressive, tous les modèles ont montré des VPN élevées, comprises entre 87 et 92% chez les filles et entre 95% et 97% chez les garçons. Les valeurs de spécificités obtenues par les modèles étaient comprises entre 74% et 85% chez les filles et entre 76% et 94% chez les garçons.

Tableau 9: Performances des modèles sur le set de validation externe

	Filles			Garçons		
	LASSO	SGD	RF	LASSO	SGD	RF
Kappa	0,3091	0,3114	0,2874	0,1981	0,2393	0,1958
Sensibilité	0,6921	0,6190	0,4571	0,7033	0,5932	0,2712
Spécificité	0,7388	0,7804	0,8521	0,7611	0,8394	0,9362
VPP	0,3482	0,3625	0,3840	0,1800	0,2161	0,2406
VPN	0,9225	0,9104	0,8862	0,9718	0,9651	0,9451

4.3.3. Variables les plus importantes

4.3.3.1. Modèles dans la population des filles

Le modèle de régression LASSO distingue parmi les 93 variables inclues, 53 variables avec un coefficient non nul. Sur ces 53 variables seulement 16 ont été identifiées comme des variables explicatives « significatives » (i.e leur permutation impacte significativement la qualité prédictive du modèle). En ce qui concerne les modèles d'agrégation d'arbres, 64 variables ont été identifiées comme des variables explicatives « significatives » du modèle SGD et 71 pour le modèle RF. La Figure 12 illustre l'importance de chacune des 93 variables explicatives dans chaque modèle et le Tableau 10 représente les 10 premières variables les plus importantes. Parmi ces dernières, quatre variables sont communes aux trois modèles : *« Actuellement, que pensez-vous de l'école ?* », *« Selon vous, la scolarité est la seule chose qui compte pour vos parents »*, *« Pour vous, est-ce important d'être mince »* et *« Manger est pour vous (un plaisir/une contrainte) »*. Notons que *« Actuellement, que pensez-vous de l'école ? »* est la première variable explicative des modèles LASSO et SGD avec une valeur d'importance d'au moins deux fois supérieure à n'importe quelle autre variable. La concordance entre les valeurs d'importance telle que mesurée par les coefficients de corrélation intra classe de cohérence est forte entre les modèles LASSO et SGD (ICC de cohérence= 0,89 IC95% [0,83 ; 0,92]) mais moindre entre les modèles SGD et RF (ICC de cohérence 0,52 IC95% [0,36 ; 0,66]) ou entre les modèles LASSO et RF (ICC de cohérence = 0,36 IC95% [0,17 ; 0,52]).

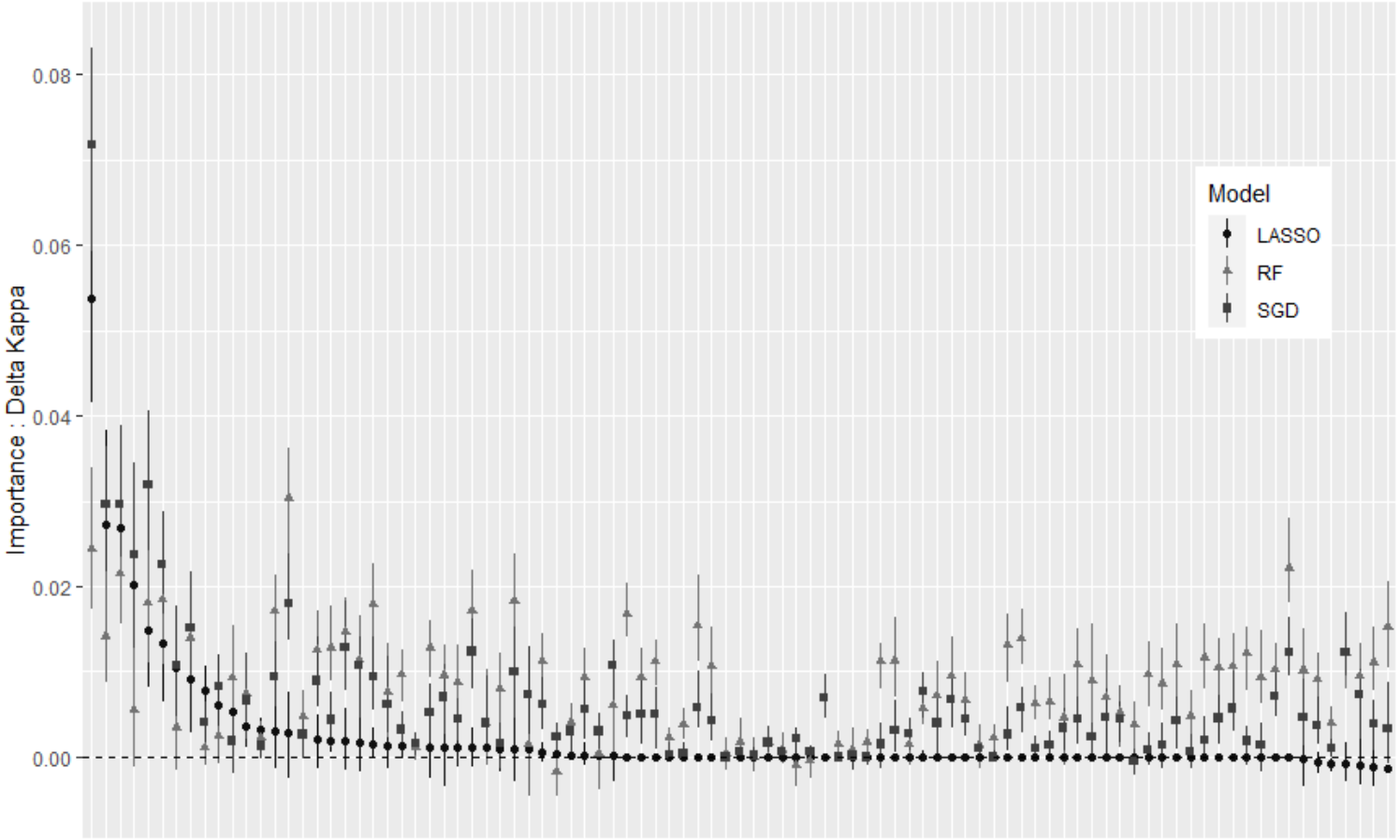
La comparaison des 10 variables les plus importantes pour chaque modèle met en lumière des groupes de variables majeures dans l'un et/ou l'autre des modèles d'agrégation d'arbres qui sont de faible importance dans les modèle LASSO *: « Combien d'amis avez-vous dans la réalité/rencontrés uniquement sur internet »*, *« le temps de jeu vidéo par jour le week-end/pendant les vacances » et « pratiquez-vous un sport régulièrement ».*

Tableau 10: Importance des dix variables impactant le plus chaque modèle dans la population des filles

	Coefficient LASSO	*Rang lasso*	*LASSO*	*Rang SGD*	*SGD*	*Rang RF*	*RF*
Actuellement que pensez-vous de l'école ?	+0,8781	**1**	**0,054[0,041;0,065]***	**1**	**0,072[0,059;0,083]***	**2**	**0,024[0,017;0,034]***
Avoir le sentiment d'être décalé (s'endormir, se réveiller très tard)	+0,6041	**2**	**0,027[0,018;0,038]***	**3**	**0,030[0,022;0,036]***	15	0,014[0,009 ; 0,022]*
La scolarité est selon vous la seule chose qui compte pour vos parents	+0,5149	**3**	**0,027[0,018;0,034]***	**4**	**0,030[0,023;0,039]***	**4**	**0,022[0,016;0,027]***
Généralement y-a-t-il des disputes dans votre famille ?	+0,5713	**4**	**0,020[0,011;0,030]***	**5**	**0,024[0,013;0,034]***	61	0,006[-0,001 ; 0,015)
Pour vous, est-ce important d'être mince	+0,4096	**5**	**0,015[0,008;0,023]***	**2**	**0,032[0,024;0,041]***	**7**	**0,018[0,011;0,024]***
Manger est pour vous ?un plaisir ou une contrainte	+0,5067	**6**	**0,013[0,006;0,020]***	**6**	**0,023[0,017;0,029]***	**5**	**0,018[0,011;0,025]***
Parlez-vous facilement avec vos parents de votre santé	-0,4245	**7**	**0,010[0,006;0,015]***	13	0,011[0,005;0,018]*	71	0,003[-0,002;0,009]
Vos amis pensent-ils du bien de vous	-0,4000	**8**	**0,009[0,003;0,015]***	**8**	**0,015[0,009;0,022]***	**16**	**0,014[0,009 ; 0,018]***
Quand vous buvez de l'alcool c'est plutôt seul	+0,7752	**9**	**0,008[0,004;0,011]***	49	0,004[0,001;0,007]*	84	0,001[-0,001;0,003]
Avez-vous déjà participé à des jeux dangereux ?	+0,3548	**10**	**0,006[0,003;0,009]***	20	0,008[0,004;0,012]*	72	0,003[-0,001;0,007]
IMC	-0,1494	14	0,003[-0,001;0,007]	17	0,009[0,006;0,015]*	**9**	**0,017[0,012;0,021]***
Combien d'amis avez-vous dans la réalité ?	-0,1272	15	0,003[- 0,003 ; 0,008]	**7**	**0,018[0,014;0,024]***	**1**	**0,030[0,022;0,036]***
Vous pesez-vous régulièrement à votre domicile ?	-0,1096	19	0,002[-0,001; 0,006]	**9**	**0,013[0,008;0,018]***	14	0,015[0,011; 0,019]*
Combien d'amis avez-vous rencontré uniquement sur internet	-0,0316	21	0,002[0,000;0,003]	18	0,009[0,006;0,014]*	**8**	**0,018[0,013;0,023]***
Temps de jeux vidéo par jour le weekend	-0,0519	28	0,001[-0,001 ; 0,003]	**10**	**0,012[0,008;0,016]***	**10**	**0,017[0,013;0,022]***
Pratiquez-vous régulièrement un sport	-0,1084	31	0,001[-0,003;0,005]	16	0,010[0,005;0,015]*	**6**	**0,018[0,011;0,024]***
Temps de jeux vidéo par jour pendant les vacances	0	46	_	11	0,012[0,009;0,017]*	**3**	**0,022[0,018;0,028]***

Légende : Les variables en gras font partie des 10 variables les plus importantes du modèle correspondant. Les variables en gras avec * impactent significativement les performances prédictives du modèle (IC95% n'incluant pas zéro).

Figure 12: Moyenne des Δkappa et écarts-type des 93 variables explicatives (représentation de l'importance de chacune des 93 variables explicatives) dans la population des filles.



Légende : Les 93 variables ont été classées dans l'ordre d'importance décroissant du modèle de régression LASSO

4.3.3.2. Modèles dans la population des garçons

La Figure 13 illustre l'importance de chacune des 93 variables explicatives dans chaque modèle. Le modèle de régression LASSO inclut 37 variables avec un coefficient non nul. Il n'y a que 3 variables qui impactent « significativement » les performances de prédiction du modèle LASSO, alors qu'il y en a 52 dans le modèle SGD et 63 dans le modèle RF. Les valeurs d'importance sont modérément concordantes entre les modèles LASSO et SGD (ICC de cohérence= 0,39 IC95% [0,21 ; 0,55]) et entre les modèles SGD et RF (ICC de cohérence 0,30 IC95% [0,11 ; 0,48]) mais aucune concordance n'a été observée entre les valeurs d'importance des modèles LASSO et RF (ICC de cohérence = 0,00 IC95% [-0,20 ; 0,20]). Comme le montre le Tableau 11, parmi les 10 variables les plus importantes, il n'y a aucune variable commune aux trois modèles. Néanmoins, on note qu'au moins une variable de consommation d'alcool (« *Quand vous buvez de l'alcool c'est plutôt seul », « Quand vous buvez de l'alcool c'est plutôt en soirées ou avec des copains », « intensité de l'ivresse la dernière fois que vous avez bu »,* ou «*consommation d'alcool au cours des 30 derniers jours)* une variable sur l'école (« *Actuellement que pensez-vous de l'école » ou « Selon vous la scolarité est la seule chose qui compte pour vos parents »)* et une variable d'alimentation et/ou IMC *(« Manger est pour vous un plaisir ou une contrainte/obligation ?», « Pour vous est ce important d'être mince », « IMC »* sont classées parmi les 10 premières variables de chacun des modèles.

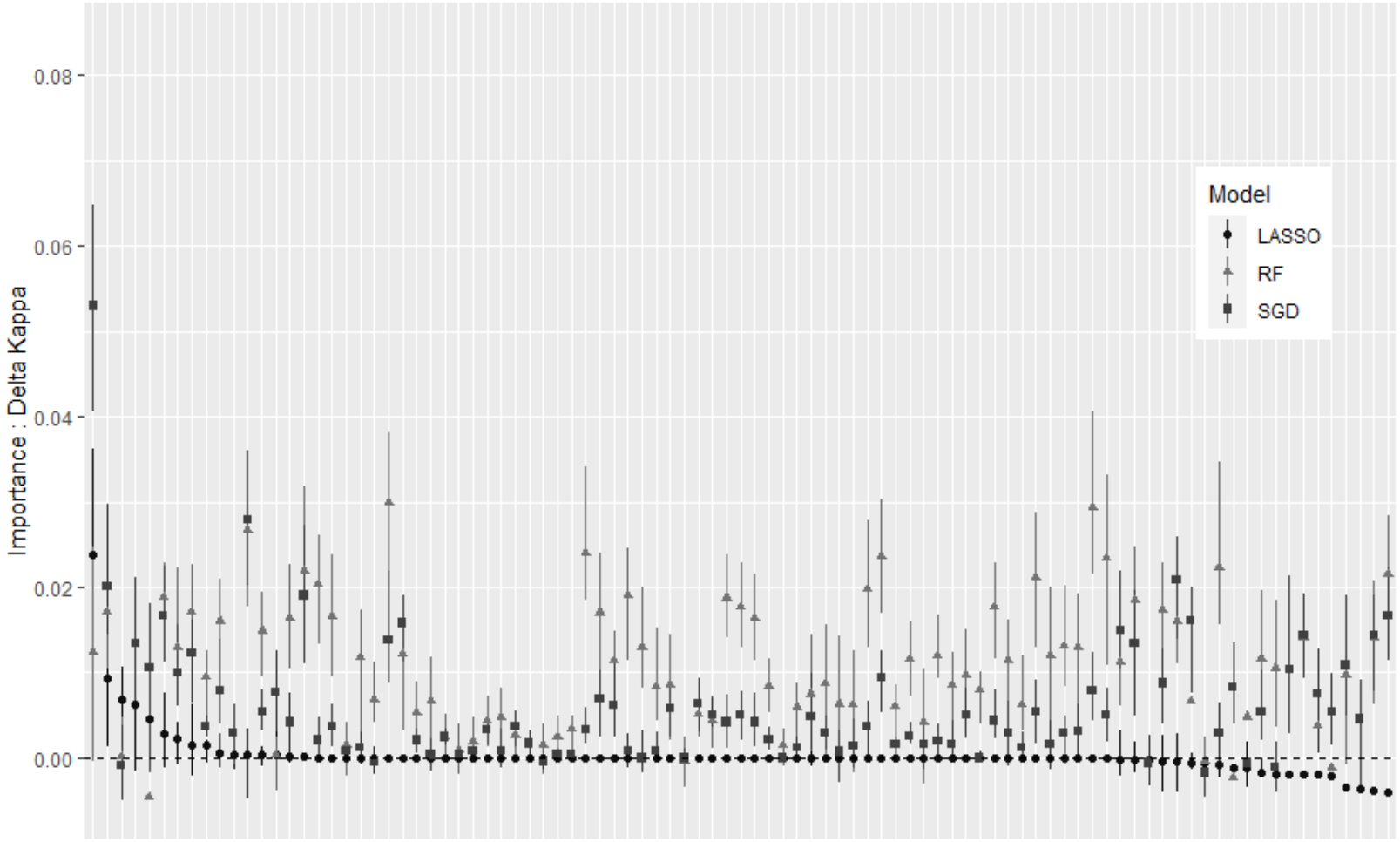
Le Tableau 11 met en lumière des groupes de variables se retrouvant dans les 10 premières variables des deux modèles d'agrégation d'arbres mais qui sont de faible importance dans le modèle LASSO des garçons : le sport (pratique d'un sport en loisirs), et le temps de jeu vidéo.

Tableau 11 : Importance des dix variables impactant le plus chaque modèle dans la population des garçons.

	Coefficients LASSO	*Rang LASSO*	*LASSO*	*Rang SGD*	*SGD*	*Rang RF*	*RF*
Actuellement que pensez-vous de l'école ?	+0,8401	**1**	**0,024[0,12;0,036]***	**1**	**0,053[0,041;0,065]***	35	0,012[0,000 ; 0,025]
Vos amis pensent-ils du bien de vous	-0,5568	**2**	**0,009[0,001;0,016]***	**4**	**0,020[0,014;0,030]***	**21**	**0,017[0,010 ; 0,026]***
Quand vous buvez de l'alcool c'est plutôt seul	+0,6616	**3**	**0,007[0,002;0,011]***	91	-0,001[-0,005 ; 0,004]	82	0,000[-0,004 ; 0,005]
D'avoir le sentiment d'être décalé (s'endormir, se réveiller très tard)	+0,5399	4	0,006[-0,002;0,014]	14	0,013[0,006;0,013]*	92	-0,009[-0,02;-0,001]
Généralement y-a-t-il des disputes dans votre famille ?	+0,5031	5	0,005[-0,006;0,010]	18	0,011[-0,002; 0,018]	87	-0,005[-0,005; 0,004]
Manger est pour vous ? (un plaisir ou une contrainte)	+0,3365	6	0,003[-0,001;0,008]	**6**	**0,017[0,011;0,023]***	14	0,019[0,012 ; 0,023]*
Enseignement général versus professionnel ou agricole	+0,0789	7	0,002[-0,001;0,004]	20	0,010[0,006; 0,016]*	32	0,013[0,006; 0,022]*
Enseignement professionnel versus général ou agricole	-0,1987	8	0,002[-0,002;0,006]	16	0,012[0,006 ; 0,016]*	20	0,017[0,010 ; 0,023]*
Attirance sexuelle (homosexuelle)	+0,1111	9	0,001[-0,001;0,002]	47	0,004[0,002 ; 0,005]*	48	0,010[0,005 ; 0,013]*
Vous regardez régulièrement dans un miroir ?	+0,0590	10	0,001[-0,001;0,003]	24	0,008[0,004 ; 0,014]*	26	0,016[0,010 ; 0,021]*
Intensité ivresse la dernière fois que vous avez bu	-0,0122	11	-0,001[-0,001 ; 0,001]	**8**	**0,016[0,008;0,020]***	58	0,007[-0,005 ; 0,015]
Temps de jeux vidéo dans la semaine	-0,1969	12	0,000[-0,005 ; 0,003]	**2**	**0,028[0,020;0,036]***	**3**	**0,027[0,018;0,035]***
Pratiquez-vous un sport régulièrement ? (non, en loisir, en compétition)	-0,0182	13	-0,001[-0,002 ; 0,001]	57	0,003[-0,001 ; 0,006]	**7**	**0,022[0,016;0,035]***
Selon vous la scolarité est la seule chose qui compte pour vos parents	+0,3830	16	0,000[-0,005 ; 0,007]	**5**	**0,019[0,011;0,027]***	**8**	**0,022[0,014;0,032]***
Loisir préféré : pratique sport	-0,2531	25	-0,004[-0,009 ; 0,000]	**7**	**0,017[0,011;0,022]***	**9**	**0,022[0,014;0,028]***
Avez-vous déjà dragué ?	+0,1404	32	0,000[-0,002 ; 0,003]	**10**	**0,015[0,010;0,022]***	44	0,011[0,006 ; 0,017]*
Pour vous est-ce important d'être mince ?	+0,0987	36	0,000[-0,004 ; 0,003]	**3**	**0,021[0,014;0,026]***	27	0,016[0,011 ; 0,020]*
Consommation régulière d'alcool les 30 derniers jours	0	38	-	13	0,014[0,009 ; 0,022]*	**1**	**0,030[0,021;0,038]***
Quand vous buvez de l'alcool c'est plutôt en soirées ou avec des copains	0	39	-	**9**	**0,016[0,011;0,019]***	36	0,012[0,003 ; 0,019]*
IMC	0	40	-	50	0,003[0,002 ; 0,006]*	**4**	**0,024[0,018;0,034]***
Combien d'amis avez-vous rencontrés uniquement sur internet	0	41	-	21	0,009[0,005 ; 0,013]*	**5**	**0,024[0,017;0,030]***
Pensez-vous connaitre vos limites dans la pratique d'un sport à risque	0	42	-	33	0,005[0,002 ; 0,009]*	**10**	**0,021[0,013;0,029]***
Temps de jeux vidéo pendant le week-end	0	43	-	25	0,008[0,004 ; 0,012]*	**2**	**0,029[0,022;0,041]***
Temps de jeux vidéo pendant les vacances	0	44	-	37	0,005[0,002 ; 0,008]*	**6**	**0,023[0,011;0,033]***

Légende : Les variables en gras font partie des 10 variables les plus importantes du modèle correspondant. Les variables en gras avec * impactent significativement les performances prédictives du modèle (IC95% n'incluant pas zéro).

Figure 13: Moyenne des Δkappa et écarts-type des 93 variables explicatives (représentation de l'importance de chacune des 93 variables explicatives) chez les garçons.



Légende : Les 93 variables explicatives ont été classées dans l'ordre d'importance décroissant du modèle de régression LASSO

4.4. Discussion

4.4.1. Rappel des objectifs et de la méthode

Dans cet axe de thèse, je me suis intéressée à l'apport des méthodes de DMML dans la modélisation du risque de présenter une forte symptomatologie dépressive. Ce, à partir d'une large gamme de potentielles variables explicatives, sociodémographiques (âge, statut scolaire, structure familiale, niveau d'éducation du père et de la mère), liées au mode de vie (consommation de substances, loisirs, sport, expérience sexuelle, sommeil), aux relations familiales et amicales et au corps (IMC, alimentation, image corporelle…). J'ai comparé ici trois modèles : deux méthodes d'agrégation d'arbres (SDG et RF) et une régression logistique pénalisée par LASSO (méthode issue de la régression logistique standard très employée en épidémiologie tout en faisant partie des méthodes de DMML). La qualité des modèles construits a été évaluée par la mesure des performances de prédiction du modèle sur des observations non utilisées lors de l'apprentissage. L'utilisation du Kappa de Cohen a permis de mesurer la performance des modèles construits par rapport à la performance qu'ils auraient obtenus par simple hasard. En plus de cette mesure principale de performance, la sensibilité et la VPP ont été calculées afin d'évaluer les capacités des modèles à trouver un équilibre entre les vrais positifs et les faux négatifs, dans un contexte de faible prévalence.

4.4.2. Qualité prédictive générale

Indépendamment de toute comparaison inter-modèle, les kappas obtenus ont montré que les qualités prédictives des modèles étaient modérées. Les VPP obtenues par ces modèles peuvent paraitre assez faibles puisqu'elles étaient au mieux de 0,38 pour les filles et 0,24 pour les garçons. Etant donné que la probabilité qu'un adolescent de la population ait une symptomatologie dépressive était de 0,17 chez les filles et 0,07 chez les garçons (prévalence présentée dans le paragraphe 3.2.4 page 65) ; ces VPP reflétaient donc une performance 2 à 3 fois (x 2,26 chez les filles et 3,48 chez les garçons) meilleure que ce à quoi on pourrait s'attendre par le fait du hasard. Ce niveau de qualité d'ajustement modèle / données était attendu, de nombreux facteurs expliquant la symptomatologie dépressive n'étant pas présents dans les données utilisées : les événements de vie traumatisant, le harcèlement, les facteurs religieux, les stratégies de régulation des émotions, les stratégies d'adaptions (30). Quoi qu'il en soit, il est important de rappeler que Kappa, sensibilité et VPP sont utilisés ici uniquement pour évaluer la qualité du modèle en mesurant sa capacité de généralisation à de « nouvelles données » et que cette étude n'a évidemment pas pour but de construire un instrument de détection ou de diagnostic.

Je n'ai trouvé aucune étude dans la littérature ayant modélisé les facteurs de risque de la dépression ou la symptomatologie dépressive chez les adolescents avec des méthodes DMML. D'une manière générale, dans le champ de l'épidémiologie en santé mentale des adolescents, les études évaluant des

méthodes DMML sont rares. Néanmoins, quelques études récentes ont utilisé ces méthodes afin de prédire d'autres évènements de santé mentale des adolescents tels que le risque de développer un trouble de déficit de l'attention/hyperactivité (111), les tentatives de suicide (112–114), des « problèmes » de santé mentale (115). Certaines études en épidémiologie clinique ont aussi utilisé des méthodes DMML par exemple afin de prédire l'efficacité spécifique, de traitements, contre la dépression (116,117). Outre des variables d'intérêts cliniques différentes, toutes ces études sauf une (112) se situaient dans une approche explicative clinique et utilisaient les antécédents de troubles mentaux comme variables explicatives du modèle soit directement par les diagnostics et/ou les traitements médicamenteux, soit avec des échelles de mesure des symptômes.

En regard des différences, comparer le niveau atteint des Kappa ou VPP de mes modèles à ceux obtenus dans la littérature est donc peu pertinent.

4.4.3. Comparaison des performances des différents modèles

Dans cette analyse, les trois modèles développés ont abouti à des coefficients Kappa similaires ; les écarts étaient minimes entre les meilleurs et les pires résultats (0,025 chez les filles et 0,044 chez les garçons). Sur le critère de sensibilité les différences sont plus marquées. En effet, le modèle LASSO généré, avec une sensibilité de l'ordre de 70% chez les filles comme chez les garçons, a montré une capacité de détection des adolescents avec une forte symptomatologie dépressive meilleure que celles des SGD, elle-même nettement meilleure que celle des RF. Réciproquement, sur le critère de la VPP, les modèles RF étaient classés en tête suivis des modèles SGD et des modèles LASSO ; ce qui reflète une perte de précision liée au gain de sensibilité. Il est donc difficile - à partir des résultats obtenus - de déterminer l'algorithme le plus pertinent quant à la modélisation de ces données. Néanmoins, on peut conclure que les méthodes d'agrégation d'arbres n'ont pas présenté de gains majeurs par rapport à la méthode LASSO dans la modélisation de la symptomatologie dépressive.

Ces conclusions sont relativement similaires à celles de l'étude de Miché et al. (113) qui avait comparé les performances prédictives d'un modèle de régression logistique standard, d'un modèle LASSO et d'un modèle RF en utilisant les données d'une cohorte de 2793 adolescents et jeunes adultes (14-24 ans à l'inclusion). Leurs modèles visaient à prédire le risque de tentatives de suicides en fonction de 16 facteurs de risque dont les plus importants se sont avérées être les antécédents de tentatives de suicides, le nombre de diagnostic de troubles mentaux (DSM-IV), la recherche préalable d'une aide pour tout type de difficulté psychologique et le niveau d'étude. Les quatre modèles de prédiction, ont montré une performance comparable mesurée *via* une aire sous la courbe ROC (AUC). Comme c'est le cas sur nos données, le modèle LASSO présentait une bien meilleure sensibilité que le modèle RF (0,212 versus 0,028), et une moindre VPP (0,716 versus 0,870). Jung et al. ont aussi obtenu des qualités

prédictives sensiblement égales (sensibilités et VPP entre 0,77 et 0,79) avec des modèles RF, Gradient Boosting, Machines à Vecteur de support, réseaux de neurones et régression logistique pour classer les adolescents ayant des antécédents d'idées/tentatives suicidaires (112). Au contraire, Walsh et al. (2018) ont conclu à la supériorité des modèles RF sur une cohorte rétrospective et longitudinale d'adolescents et de témoins, en utilisant les données du dossier médical électronique ; le modèle RF donnant des AUC de plus de 0,8, tandis que la régression logistique donnait des AUC inférieures à 0,7 (114).

Des comparaisons entre méthodes d'agrégation d'arbres et modèles de régression (standards ou pénalisés) ont été fréquemment évaluées dans d'autres champs de la littérature biomédicale. A titre d'exemple, Olson et al. (118) ont comparé régression logistique, méthodes d'agrégation d'arbres et d'autres méthodes de DMML (dont les machines à support de vecteur ou des classifieurs naïfs de Bayes) sur une collection de 165 jeux de données publics. Ces jeux de données bien que non limités au champ biomédical incluaient de nombreux problèmes de classification de diagnostic des maladies, et de données d'association pangénomique. Globalement, les SGD et les RF se sont révélés être les deux modèles les plus performants en termes de justesse de prédiction (accuracy). Leurs performances prédictives outrepassaient notamment les résultats des régressions logistiques sur respectivement 78% et 71% des jeux de données. Notons toutefois que la régression logistique était plus performante que les SGD et les RF dans respectivement 5% et 10% des jeux de données.

Néanmoins la littérature est loin d'être consensuelle. A titre de contre-exemple, une revue systématique récente est arrivée à la conclusion opposée (119). L'objectif de cette revue était d'évaluer les performances des algorithmes DMML dans le cadre de modèles de prédiction diagnostique ou pronostique couvrant de nombreux domaines de recherche épidémiologique, tels que la psychiatrie, la cardiologie ou l'oncologie. Sur 71 articles publiés en 2016 et 2017, ils ont identifié 282 comparaisons entre un modèle issu d'une méthode DMML et un modèle de régression logistique (standard ou pénalisée). Les modèles RF (39% des articles), et SGD (23% des articles) faisaient partie des méthodes de DMML les plus fréquemment retrouvées avec les réseaux neuronaux artificiels et les SVM. Les résultats de leur méta-régression, incluant uniquement les articles qui présentaient un faible risque de biais, ont montré que les performances prédictives des modèles de régression logistique et ceux issus de méthodes DMML étaient similaires. Les performances de l'apprentissage automatique étaient plus élevées pour les articles qui présentaient un risque élevé de biais.

Mon hypothèse de départ était que les algorithmes de DMML pourraient montrer de meilleures performances prédictives que les modèles de régression logistique grâce à leur capacité intrinsèque à modéliser des interactions d'ordre élevée et à traiter le problème du « sur-ajustement ». La question

se pose donc de comprendre pourquoi certaines études dont la mienne font état de peu de gains de performance via les modèles de DMML.

Plusieurs explications ont été évoquées dans la littérature. Les performances des algorithme DMML seraient gourmandes en données et dépendraient notamment de la taille de l'échantillon et du rapport nombres d'événements par variable explicative (120). Les algorithmes DMML auraient aussi tendance à mieux fonctionner sur les jeux de données présentant un fort rapport signal/bruit (119). Enfin, leurs performances dépendraient de la complexité en haute dimension (par exemple, associations non linéaires, interactions d'ordre élevé) effectivement présente dans les données (113).

Les caractéristiques structurales des jeux de données moins favorables aux algorithmes DMML complexes sont relativement fréquentes aux jeux de données dans le champ de la santé publique et peuvent effectivement s'appliquer à notre échantillon. Ainsi de nombreuses études inclues dans la revue de Christodoulou et al., incluaient un faible nombre d'événements par prédicteur : à savoir une médiane de 8,0 événements par prédicteur dans le set d'apprentissage (de 0,3 à 6697) et une taille médiane de l'échantillon de 1250 sujets (de 72-3 994 872) (119). Dans notre étude, le rapport événements/variables était de 13,6 chez les filles et 5,1 chez les garçons. De même, le rapport signal/bruit de nos données est vraisemblablement faible, puisque la majorité des domaines explorés (sommeil, expérimentation de substances psychoactives, utilisation de jeux vidéo, etc.) sont fréquents dans la population adolescente. Il est aussi possible que la complexité en haute dimension des interactions entre facteurs de risque de la dépression ne soit pas accessible dans nos données, où les variables sont codées en catégoriel et abordées *via* des proxis simples (e.g l'utilisation des écrans est uniquement investigué *via* le temps passé à jouer à des jeux vidéo sans mesure de l'impact sur la vie de l'adolescent). Par conséquent, les résultats des modèles d'agrégation d'arbres pourraient être expliqués partiellement par les critères susmentionnés. Toutefois, il existe très probablement d'autres caractéristiques structurelles qui favorisent ou non les performances des algorithmes DMML complexes (113).

4.4.4. Conclusion sur la comparaison de la qualité des modèles

Les modèles d'agrégation d'arbres n'ont pas montré de gains majeurs par rapport à une régression pénalisée LASSO quant au classement des adolescents présentant une forte symptomatologie dépressive à partir de variables explicatives sociodémographiques, comportementales et de facteurs de stress. Les performances des modèles DMML étant fortement impactées par la structure des données, cette conclusion n'est donc absolument pas généralisable à d'autres données. Au final, de nombreux auteurs estiment qu'il n'est pas possible de prédire à l'avance quel type d'algorithme sera plus efficace sur tel type de données et que l'approche la plus efficace est de les tester pour choisir a posteriori celui qui est le plus approprié. Les méthodes de DMML se conforment au Théorème du « No free lunch » de Wolpert, dans le champ biomédical comme ailleurs, aucun algorithme n'est meilleur que tous les autres sur l'ensemble de tous les problèmes possibles (121).

4.4.5. Importance des variables

L'objectif secondaire de cet axe, consistait en la sélection des variables explicatives majeures dans la prédiction de la symptomatologie dépressive parmi les trois modèles. Pour rappel, une variable a été considérée « majeure » si la permutation de ses valeurs augmentait l'erreur du modèle (IC95% strictement positif).

Sur ce critère, entre 16 et 71 variables sur les 93 ont été identifiés selon les modèles comme importantes chez les filles pour la classification des sujets avec une forte symptomatologie dépressive ; entre 3 et 63 dans l'échantillon des garçons. Si le nombre de variables statistiquement importantes dans les modèles variait, la distribution des valeurs d'importance (Figure 12, Figure 13) montrait néanmoins des valeurs d'importance globalement faibles. En d'autres termes, individuellement leur permutation entrainait une dégradation minime des performances prédictives du modèle. Quelques variables ont toutefois montré un impact plus conséquent sur le modèle, les valeurs d'importance pouvant atteindre 0,07 pour le SGD, 0,05 pour le LASSO, et 0,03 pour les modèles RF.

Les valeurs d'importance étaient concordantes entre les modèles LASSO et SGD en particuliers chez les filles (ICC cohérence LASSO / SGD 0,89 chez les filles). Cependant les concordances étaient plus faibles entre les autres modèles, voire nulles entre les modèles LASSO et RF chez les garçons (ICC cohérence LASSO / RF 0,0 chez les garçons). Des variabilités considérables entre les valeurs d'importance de différents modèles ont été également rapportées dans d'autres études (113,122). Ceci est intéressant, en particulier dans le contexte de modèles qui ont des performances globales similaires, et corrobore l'idée de les combiner pour saisir différents niveaux d'informations sur les facteurs de risque. Il est important de rappeler que contrairement aux coefficients des modèles de

régression, les estimations de l'importance des variables dans les modèles d'agrégation d'arbres ne représentent pas la contribution unique des variables explicatives en supposant toutes les autres variables égales par ailleurs. Cette mesure quantifie plutôt l'impact marginal d'un prédicteur, en tenant compte des interactions avec tous les autres prédicteurs du modèle. Les estimations de l'importance des variables peuvent donc varier considérablement en fonction des autres variables et de la manière dont elles ont été modélisées. À titre d'exemple, chez les jeunes garçons, il existe un lien significatif entre la symptomatologie dépressive et le fait d'avoir déjà eu des rapports sexuels (42). Au contraire, chez les garçons plus âgés, la symptomatologie dépressive ont été associés au fait de ne pas avoir eu de rapports sexuels (42).

La liste des dix variables les plus importantes différait partiellement d'un algorithme à l'autre. La comparaison de ces listes a mis en lumière des points communs entre les modèles d'agrégation d'arbres (7 /10 variables en commun entre les modèles RF et SGD chez les filles, 3/10 chez les garçons). Parmi ces variables majeures dans les deux modèles d'agrégation d'arbres, certaines sont de moindre importance dans les modèles LASSO*: le nombre d'amis dans la réalité, se peser régulièrement à domicile, le temps de jeu vidéo, avoir comme loisir régulier l'activité physique et le ressenti que la scolarité est la seule chose qui compte pour ses parents.*

Sachant que les valeurs d'importance incluent les effets d'interaction, et que les interactions n'ont pas été modélisées dans les régressions LASSO, la comparaison de l'importance des variables dans différents modèles peut aider à découvrir les interactions entre les variables explicatives (clinique et statistique). Dans ce cadre, nos résultats sont cohérents avec une revue de la littérature récente de Zink et al., qui suggère des modulations de l'association temps passé sur les écrans et symptômes dépressifs par l'activité physique, et les relations aux pairs avec notamment la perception de la qualité des relations amicales (123). Ces résultats seront repris et discutés dans la discussion générale de la thèse.

Toutefois, bien d'autres mécanismes sous-jacents peuvent expliquer les différences entre modèles. Certaines études ont suggéré que les méthodes d'estimation de l'importance des variables par Permutation/Prédiction pouvaient surestimer l'importance des prédicteurs fortement associés entre eux (124). Les biais précis que les méthodes de permutation produisent dépendraient de la méthode d'apprentissage utilisée et seraient plus forts pour les méthodes flexibles telles que les forêts aléatoires ou les réseaux de neurones (125). Or, certaines des variables explicatives correspondantes font aussi partie de groupes de variables très associées (e.g temps de jeu vidéo en semaine, en week-end et en vacances, avoir comme loisir régulier l'activité physique ou le sport, et les questions spécifiques sur la pratique d'un sport en compétition). Il est quasi impossible de s'appuyer sur la littérature pour étayer

ou invalider ces résultats car dans la plupart des études utilisant des méthodes DMML, l'objectif est avant tout la création d'un outil d'aide au diagnostic ou au pronostic ; les estimateurs d'importance des variables sont très rarement comparés entre eux.

4.4.6. Forces et limites méthodologiques

Les points forts de cette étude comprennent l'analyse simultanée d'une grande variété de facteurs associés à la dépression de l'adolescent. Une étape de validation croisée imbriquée répétée a été utilisée couplée à une validation externe ce qui rend l'estimation de la qualité des modèles et du degré de sur-ajustement particulièrement robuste.

Parmi les limites de cette étude, le rapport entre nombre d'événements et nombre de prédicteurs n'a probablement pas permis une exploitation optimum des modèles issus de méthodes DMML. L'estimation de l'importance des variables était potentiellement biaisée et ces estimateurs sont débattus dans la littérature (125). Néanmoins la méthode utilisée pour estimer l'importance des variables a l'avantage d'être une méthode agnostique, ce qui permet de comparer les résultats de modèles conceptuellement très différents. Ce pourquoi, c'est une méthode populaire (126) qui a été utilisée dans des domaines variés (122,127)). Enfin, l'utilisation de la comparaison de l'importance des variables dans les différents modèles pour repérer les interactions entre les variables explicatives est questionnable dans un contexte où les performances des modèles d'agrégation d'arbres n'ont pas montré de réel gain par rapport à la régression pénalisée LASSO.

Le domaine de l'interprétation des méthodes DMML a vraiment décollé vers 2015, et est en pleine extension (126). Des recherches visent à quantifier l'incertitude des valeurs d'importance, et à les adapter aux problèmes de tests multiples. D'autres visent à éliminer les biais dus aux corrélations fortes entre prédicteurs, en utilisant un schéma de permutation conditionnelle qui respecte la distribution conjointe des données. Des approches ont été récemment implémentées pour identifier et mesurer les interactions entre variables (128–130). Ces approches sont prometteuses et pourront ultérieurement être utilisées pour compléter l'interprétation des modèles développés ici.

En conclusion, malgré les développements récents, les algorithmes d'agrégation d'arbres restent partiellement des « boîtes noires » (113). Le défi à relever consiste à établir des bonnes pratiques quant à l'interprétation des modèles DMML, en particulier, la manière de quantifier l'interprétabilité d'un modèle ou la justesse d'une interprétation du modèle (126). Certaines personnes affirment qu'il n'existe, en général, aucun modèle DMML intrinsèquement interprétable et qu'il serait même dangereux d'avoir une illusion d'interprétabilité (131). Il ne s'agit pas seulement d'un challenge mathématique ou algorithmique. Ainsi, même si l'importance des interactions entre prédicteurs peut être extraite d'un modèle, l'algorithme peut avoir utilisé les prédicteurs ininterprétable pour le cerveau

humain telles que des interactions du 10ème ordre (113). Une approche très pragmatique permettant de contourner ces difficultés consiste en l'utilisation de méthodes de partitionnement supervisée.

4.4.7. Résultats « cliniques »

Dans ce chapitre, je vais résumer les résultats marquants sur le plan de la symptomatologie dépressive. Certains de ces résultats seront comparés à la littérature et discuté dans le chapitre de « Discussion générale » (chapitre 6) de la thèse simultanément avec les résultats du chapitre 5. La variable interrogeant les sentiments vis-à-vis de l'école *(« Actuellement que pensez-vous de l'école ? »)* était la plus importante des modèles LASSO et SGD chez les filles comme chez les garçons, avec une valeur d'importance au moins deux fois supérieure à n'importe quelle autre variable.

Outre le rapport à l'école, la liste des domaines retrouvés dans les dix variables les plus importantes, tous modèles confondus, était très éclectique incluant une large palette de facteurs de risque identifiés à la fois chez les filles et les garçons :

- Le retard de phase de sommeil (avoir le sentiment d'être décalé, s'endormir très tard, se réveiller très tard)
- Les relations aux amis (nombre d'amis rencontrés sur internet, « vos amis pensent du bien de vous », et le nombre d'amis dans la réalité chez les filles)
- Les relations intrafamiliales (fréquence des disputes dans la famille + chez les filles parler facilement de sa santé avec ses parents)
- L'IMC, le rapport à l'alimentation et l'image corporelle (importance d'être mince, manger est une contrainte + chez les filles se peser régulièrement à domicile chez les filles + chez les garçons se regarder régulièrement dans un miroir)
- L'activité physique et le sport : (Pratiquer régulièrement un sport en loisir ou en compétition + uniquement chez les garçons connaitre ses limites dans les sports à risque)
- Le temps de jeu vidéo par jour (en week-end et en vacances)
- La consommation d'alcool : (boire de l'alcool plutôt seul + chez les garçons boire de l'alcool plutôt en soirée avec des copains, la fréquence de la consommation d'alcool sur le mois et l'intensité de l'ivresse)

Par ailleurs, la participation à des jeux dangereux est apparu en tant que variable majeure chez les filles, et les relations sexuelles / amoureuses/ de séduction (attirance sexuelle et avoir déjà « dragué ») chez les garçons.

La comparaison de l'importance des variables dans les différents modèles suggère des interactions entre le temps passé sur des jeux vidéo, l'activité physique, le nombre d'amis sur internet et la consommation d'alcool.

Suite aux résultats obtenus et à la difficulté d'interprétation des associations entre les variables explicatives et la symptomatologie dépressive, il est nécessaire d'étudier le rôle conjoint de plusieurs variables explicatives sur le risque de présenter une forte symptomatologie dépressive. Le rôle conjoint de ces variables, peut être évaluée par la création de différents sous-groupes d'adolescents (ou profil) à risque de présenter une forte symptomatologie dépressive. Les variables qui expliquent le contraste entre les sous-groupes à forte probabilité peuvent être mises en évidence, ce qui renforce l'interprétabilité du rôle conjoint de ces variables et des sous-groupes.

5. Intérêt d'une méthode DMML de partitionnement à la reconnaissance des adolescents à forte symptomatologie dépressive

Afin d'explorer les associations entre le profil des adolescents et la symptomatologie dépressive, une méthode supervisée de partitionnement des données a été utilisée : « Bayesian Profile Regression », appelée dans la suite de ce travail « Régression sur profil ». Il s'agit d'une méthode semi-paramétrique bayésienne, spécifiquement créée afin d'étudier l'effet joint d'un vaste ensemble de facteurs de risque sur une variable d'intérêt clinique (75,132,133). L'approche utilisée s'appuie donc sur un modèle de DMML bayésien guidée par les données (134).

5.1. Matériel et méthode

Comme précédemment, les données ont été extraites de l'étude « Processus d'adolescence ». Cette enquête interrogeait 14318 adolescents sur de nombreux sujets (paragraphe 3.2.1 page 60), notamment sur leur symptomatologie dépressive *via* l'ADRS (1263 filles qui présentaient une forte symptomatologie dépressive sur 7513 filles et 473 garçons sur 6805) dont 93 variables incluant des renseignements sociodémographiques (âge, sexe, établissement scolaire) et un ensemble d'items interrogeant des domaines connus dans la littérature pour être des marqueurs de risque ou facteurs de risque de la dépression de l'adolescent.

L'approche est résumée dans la Figure 14. Les données des filles et des garçons ont été modélisées séparément.

5.1.1. Modèle de régression sur profil

Le modèle de régression sur profil est un modèle de mélange conçu pour regrouper des sujets en clusters de manière flexible, le processus de partitionnement étant guidé à la fois par les variables explicatives et la variable d'intérêt clinique. Pour cela, le modèle réunit deux sous-modèles, ajustés conjointement et qui s'informent mutuellement (132):

1) Un « sous-modèle d'attribution » qui regroupe les individus selon leur profil de covariables explicatives. Il permet d'évaluer la probabilité qu'un sujet soit affecté à un cluster particulier. Ce modèle de mélange intègre un processus de Dirichlet. Dans ce sous-modèle, la classification est non supervisée (la variable d'intérêt clinique n'est pas utilisée ici).
2) Un « sous-modèle de maladie », qui lie les clusters à la variable d'intérêt clinique *via* un modèle de régression quantifiant ainsi le lien entre un profil de marqueurs/facteurs de risque et le risque associé pour chaque cluster.

Les deux sous-modèles sont joints par un algorithme de Monte-Carlo par Chaînes de Markov (MCMC) de sorte que l'allocation d'un individu à un cluster va dépendre à la fois des covariables dans le premier sous-modèle, et de la variable d'intérêt clinique dans le second sous-modèle.

En pratique, l'algorithme MCMC effectue de nombreuses itérations et détermine une nouvelle partition des sujets à chacune d'elle. Le nombre de clusters n'est pas défini *a priori*, mais est déterminé par les données au fil de l'algorithme, à l'inverse de certaines autres méthodes de clustering (e.g K means …). Ce nombre de clusters peut donc varier d'une itération à l'autre.

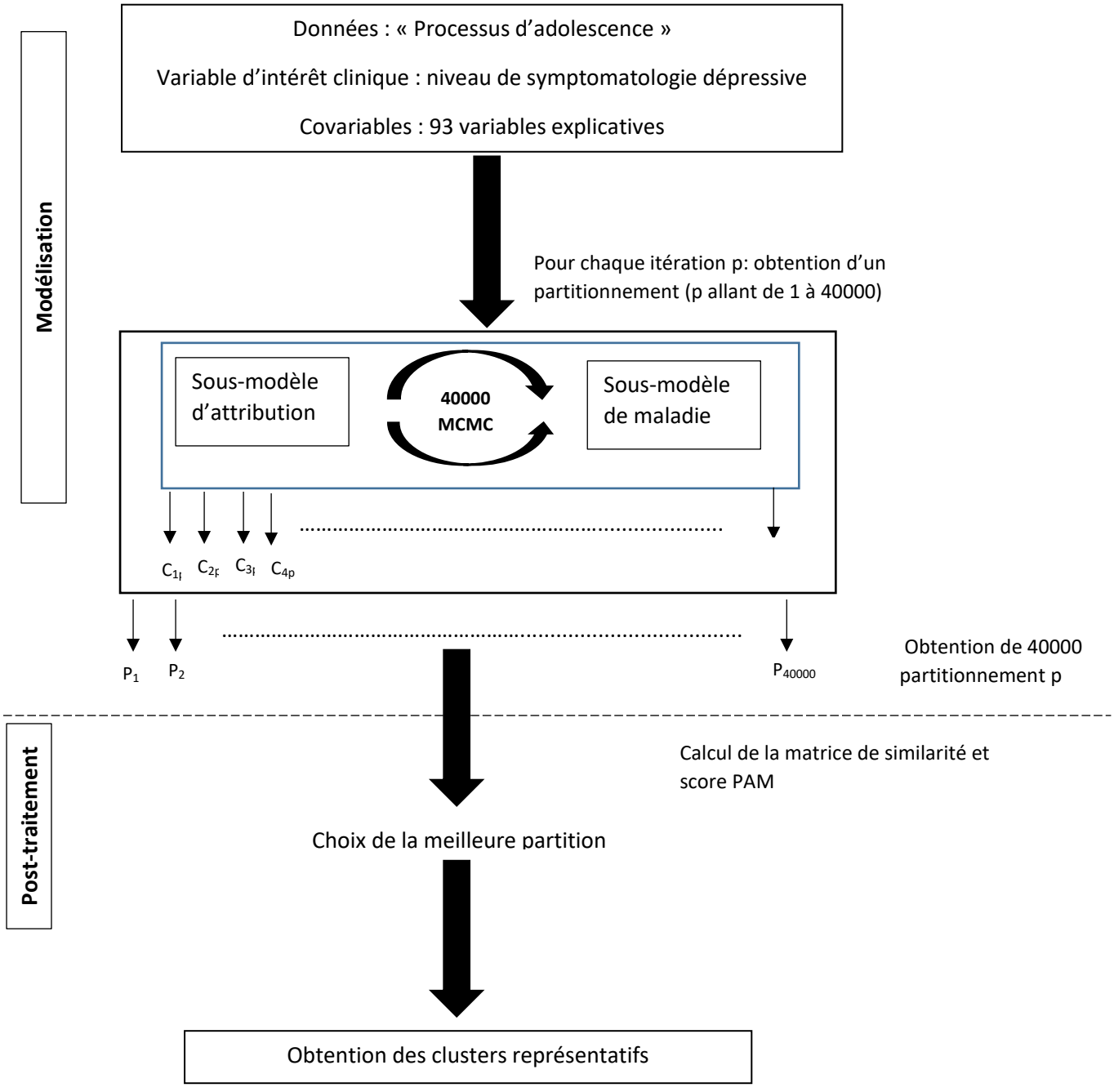
La librairie « PReMiuM » spécifiquement créé sous R par Liverani et al. en 2013 (135), pour implémenter les régressions sur profil, a été utilisé dans le cadre de cette analyse. Le nombre d'itération du MCMC a été fixé à 40000 (dont 20000 de période de rodage).

La dernière étape a consisté à sélectionner une partition optimale parmi toutes celles explorées au cours des 40000 itérations afin de synthétiser les résultats du modèle. Le choix de cette partition, s'est fait sur un critère de représentativité estimé *a posteriori* à partir d'une matrice de similarité dans laquelle l'élément (i,j) décompte le nombre de fois où les individus i et j sont dans le même cluster. La procédure mise en œuvre dans la librairie utilise l'algorithme de « Partitionnement Autour de Médoïdes » (PAM) sur cette matrice. Dans un premier temps, la meilleure partition PAM est sélectionnée pour chaque nombre possible de clusters puis un cluster représentatif final est choisi en maximisant le coefficient silhouette (i.e mesure de qualité d'une partition évaluant la dissimilarité intra et inter clusters). Dans cette analyse, afin d'obtenir des profils d'adolescents ayant des fréquences

relativement importantes, j'ai choisi d'imposer un critère supplémentaire dans le choix du partitionnement final à savoir un nombre de sujets par cluster au moins égal à 250 (~ 3,5% des sujets).

En termes statistiques, l'originalité de cette approche est double i) il s'agit d'une partition supervisée par la variable d'intérêt clinique, ii) l'unité d'inférence de la régression n'est pas le sujet mais le profil de covariables qui lui correspond.

Figure 14: Schéma explicatif du procédé de la "régression sur profil"



5.1.2. Description du partitionnement sélectionné

Dans le cadre de ce travail de thèse, le modèle de « régression sur profil » a permis de construire une partition en clusters définis à la fois par les probabilités des valeurs de 93 covariables et par le niveau de symptomatologie dépressive. La description de cette partition aboutit à l'estimation du risque de présenter une forte symptomatologie dépressive associée à chaque cluster, l'identification des clusters significativement « à haut risque » et enfin la caractérisation en termes de profils de variables explicatives des clusters identifiés.

En pratique, la première étape a consisté à décrire le nombre de clusters dans la partition finale puis le taux de sujets présentant une forte symptomatologie dépressive dans chaque cluster. Les clusters à « haut risque » ont été déterminés par un modèle de régression logistique prédisant le niveau de symptomatologie dépressive en fonction de l'assignation aux clusters. Les distributions en termes d'âge, de région, et d'établissement scolaire au sein de la partition ont été décrites systématiquement de façon à caractériser socio-démographiquement la population majoritaire de chaque cluster.

5.1.2.1. Caractérisation des différents profils de cluster à « haut risque »

La caractérisation en terme de profils de variables explicatives des clusters implique de fait une étape de sélection des variables.

En effet, les clusters sont créés en fonction d'un vaste ensemble de variables explicatives. Il est probable qu'un nombre non négligeable d'entre elles ne contribue que peu, voire pas du tout à ce partitionnement. Afin de caractériser les clusters, il est indispensable de s'affranchir des variables explicatives les moins discriminantes. Cela peut être formulé comme une question de sélection de variables, sélection d'autant plus utile que le nombre de variables à étudier est élevé. Différentes approches ont été proposées (135). J'ai utilisé l'option de sélection des variables « Continue » de la librairie « PReMiuM » (135) qui implémente la formulation proposée par Papathomas et al. (136). Pour chaque covariable, est généré un « poids de sélection latent », qui s'interprète comme un *proxy* de l'importance de la variable dans le partitionnement. Ce paramètre varie entre 0 et 1 ; des valeurs proches de 1 suggèrent une forte probabilité de contribuer au partitionnement, tandis que des valeurs plus proches de 0 suggèrent une moindre contribution à la formation des clusters. La distribution de ce poids de sélection médian peut être utilisée afin de sélectionner les variables qui ont contribué plus que les autres au modèle (136). Pour cette analyse, les variables avec un poids de sélection au moins égal à 0,95 ont été retenues pour la caractérisation des profils.

5.2. Résultats

5.2.1. Répartition des adolescents dans les clusters : partition finale

Trente-quatre clusters ont été définis par le modèle de régression sur profil : 20 clusters pour les filles et 14 clusters pour les garçons. Le Tableau 12 pour les filles et Tableau 13 pour les garçons présentent les caractéristiques sociodémographiques des 34 clusters. On remarque que le partitionnement des sujets dans les clusters a globalement regroupé les adolescents scolarisés dans le même type d'établissement (collège ou lycée général et technologique ou lycée professionnel ou agricole selon le cluster).

Ainsi :

- 12 clusters regroupaient entre 52,44% et 100% de collégiens
- 12 clusters regroupaient entre 79,85% et 99,3% d'adolescents en lycée général et technologique.
- 10 clusters regroupaient entre 93,47% et 100% d'adolescents respectivement scolarisés en lycée professionnel ou agricole.

Dans la population des filles, la prévalence de forte symptomatologie dépressive variait de 2,03% dans le cluster 1 à 44,77% dans le cluster 20 (moyenne inter-cluster 16,87%). Dans la population des garçons, la prévalence de forte symptomatologie dépressive, variait de 1,75% dans le cluster 1 à 16,00% dans le cluster 14 (moyenne inter-cluster 7,21% pour les garçons).

Les modèles de régression logistique ont permis d'identifier quatre clusters chez les filles et cinq chez les garçons avec des risques de forte symptomatologie dépressive significativement plus élevés que la moyenne des clusters. Ces clusters (i.e clusters 17 à 20 chez les filles et 10 à 14 chez les garçons) ont été labellisés par la suite respectivement comme clusters à « haut risque ».

Le Tableau 14, présente les prévalences de forte symptomatologie dépressive pour chacun des clusters et les résultats de la régression logistique permettant d'identifier les clusters à « haut risque » selon le genre

Tableau 12 : Description de tous les clusters obtenus dans l'échantillon des filles en fonction de l'âge et du type d'établissement scolaire

Clusters N	Cluster 1 541	Cluster 2 302	Cluster 3 563	Cluster 4 328	Cluster 5 471	Cluster 6 259	Cluster 7 300	Cluster 8 339	Cluster 9 289	Cluster 10 311
Région										
Provence-Alpes-Côte d'Azur	**330(61,00%)**	**151(50,00%)**	**335(59,50%)**	**182(55,49%)**	**233(49,47%)**	48(18,53%)	95(31,67%)	81(23,89%)	67(23,18%)	**201(64,63%)**
Nouvelle Aquitaine	6(1,11%)	56(18,54%)	63(11,19%)	59(17,99%)	32(6,79%)	**161(62,16%)**	26(8,67%)	**240(70,80%)**	**203(70,24%)**	74(23,79%)
Ile de France	205(37,89%)	95(31,46%)	165(29,31%)	87(26,52%)	206(43,74%)	50(19,31%)	**179(59,66%)**	18(5,31%)	19(6,58%)	36(11,58%)
Age moyen(E-T)	13,36 (0,48)	15,69 (0,89)	16,16 (0,91)	16,05 (0,97)	13,71 (0,73)	16,27 (1,13)	14,08 (1,15)	17,05 (1,32)	16,74 (1,21)	15,87 (1,11)
Age en catégorie										
entre 13 et 15 ans	**541(100,00%)**	3(1,00%)	6(1,07%)	0(0,00%)	**418(88,75%)**	2(0,77%)	**223(74,34%)**	1(0,29%)	0(0,00%)	30(9,64%)
Entre 15 et 18 ans	0(0,00%)	**287(95,03%)**	**533(94,67%)**	**309(94,21%)**	53(11,25%)	**226(87,26%)**	73(24,33%)	**210(61,95%)**	**209(72,32%)**	**263(84,57%)**
Plus de 18ans	0(0,00%)	12(3,97%)	24(4,26%)	19(5,79%)	0(0,00%)	31(11,97%)	4(1,33%)	128(37,76%)	80(31,48%)	18(5,79%)
Type d'établissement										
Collège	**539(100,00%)**	26(8,64%)	4(0,71%)	9(2,77%)	**451(96,16%)**	3(1,17%)	**251(85,67%)**	0(0,00%)	0(0,00%)	35(11,29%)
Lycée général et technologique	0(0,00%)	**275(91,36%)**	**557(99,29%**	**313(96,31%)**	18(3,84%)	1(0,39%)	42(14,33%)	5(1,48%)	4(1,39%)	**275(88,71%)**
Lycée professionnel et agricole	0(0,00%)	0(0,00%)	0(0,00%)	3(0,92%)	0(0,00%)	**251(98,44%)**	0(0,00%)	**333(98,52%)**	**283(98,61%)**	0(0,00%)

Clusters N	Cluster 11 309	Cluster 12 264	Cluster 13 317	Cluster 14 521	Cluster 15 333	Cluster 16 391	Cluster 17* 594	Cluster 18* 393	Cluster 19* 411	Cluster 20* 277
Région										
Provence-Alpes-Côte d'Azur	53(17,15%)	**110(41,67%)**	94(29,65%)	77(14,78%)	**221(66,37%)**	**251(64,19%)**	**387(65,15%)**	**213(54,20%)**	115(27,98%)	**165(59,57%)**
Nouvelle Aquitaine	**152(49,19%)**	63(23,86%)	22(6,94%)	**382(73,32%)**	10(3,00%)	15(3,84%)	71(11,95%)	66(16,79%)	**261(63,50%)**	49(17,69%)
Ile de France	104(33,66%)	91(34,47%)	**201(63,41%)**	62(11,90%)	102(30,63%)	125(31,97%)	136(22,90%)	114(20,01%)	35(8,52%)	63(22,74%)
Age moyen(E-T)	16,41(1,20)	16,13(1,13)	16,03(1,06)	16,98(1,21)	13,58(0,50)	13,63(0,82)	16,13(1,11)	16,32(1,31)	16,68(1,23)	14,15(1,05)
Age en catégorie										
entre 13 et 15 ans	3(0,97%)	6(2,27%)	2(0,63%)	0(0,00%)	**332(99,70%)**	**368(94,11%)**	39(6,56%)	30(7,64%)	5(1,22%)	**198(71,48%)**
Entre 15 et 18 ans	**258(83,50%)**	**222(84,09%)**	**285(89,61%)**	**357(68,52%)**	1(0,30%)	20(5,12%	**500(84,18%)**	**288(73,28%)**	**313(76,15%)**	74(26,71%)
Plus de 18ans	48(15,53%)	36(13,64%)	30(9,46%)	164(31,48%)	0(0,00%)	3(0,77%)	55(9,26%)	75(19,08%)	93(22,63%)	5(1,81%)
Type d'établissement										
Collège	4(1,33%)	29(11,11 %)	17(5,38%)	1(0,19%)	**327(98,79%)**	**360(95,11%)**	58(9,80%)	77(19,64%)	4(0,99%)	**240(89,89%)**
Lycée général et technologique	0(0,00%)	**218(83,53%)**	**299(94,62%)**	0(0,00%)	4(1,11%)	13(3,34%)	**533(90,03%)**	**313(79,85%)**	0(0,00%)	25(9,36%)
Lycée professionnel et agricole	**296(98,67%)**	14(5,36%)	0(0,00%)	**512(99,81%)**	0(0,00%)	6(0,15%)	1(0,17%)	2(0,51%)	**400(99,01%)**	2(0,75%)

Légende : **En gras,** type d'établissement le plus représenté dans chaque cluster, * Cluster identifié à haut risque de symptomatologie dépressive

Tableau 13: Description de tous les clusters obtenus dans l'échantillon des garçons en fonction de l'âge et du type d'établissement scolaire

Clusters	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**	**Cluster 7**
N	628	512	377	612	414	623	515
Région							
Provence-Alpes-Côte d'Azur	235(37,42%)	**298(58,20%)**	**264(70,03%)**	**345(56,37%)**	**168(40,58%)**	**341(54,74%)**	**293(56,89%)**
Nouvelle Aquitaine	**370(58,92%)**	33(6,44%)	25(6,63%)	158(25,82%)	109(26,33%)	120(19,26%)	24(4,66%)
Ile de France	23(3,66%)	181(35,36%)	88(23,34%)	109(17,81%)	137(33,09%)	162(26,00%)	198(38,45%)
Age moyen(E-T)	16,57(1,17)	13,45(0,51)	13,55(0,50)	16,09(1,02)	15,80(0,94)	16,09(1,02)	13,51(0,54)
Age en catégorie							
entre 13 et 15 ans	2(0,32%)	**510(99,61%)**	**376(99,73%)**	4(0,65%)	3(0,73%)	2(1,32%)	**506(98,25%)**
Entre 15 et 18 ans	**500(79,62%)**	2(0,39%)	1(0,27%	**558(91,18%)**	**390(94,20%)**	**566(90,85%)**	9(1,75%)
Plus de 18ans	126(20,06%)	0(0,00%)	0(0,00%)	50(8,17%)	21(5,07%)	55(8,83%)	0(0,00%)
Type d'établissement							
Collège	0(0,00%)	**504(100,00%)**	**373(100%)**	43(7,14%)	29(7,11%)	2(0,32%)	**503(98,02%)**
Lycée général et technologique	0(0,00%)	0(0,00%)	0(0,00%)	**534(88,70%)**	**338(82,84%)**	**612(98,55%)**	5(0,98%)
Lycée professionnel et agricole	**617(100,00%)**	0(0,00%)	0(0,00%)	25(4,16%)	41(10,05%)	7(1,13%)	0(0,00%)

Clusters	**Cluster 8**	**Cluster 9**	**Cluster 10***	**Cluster 11***	**Cluster 12***	**Cluster 13***	**Cluster 14***
N	342	438	462	604	490	488	300
Région							
Provence-Alpes-Côte d'Azur	99(28,95%)	124(28,31%)	**195(42,21%)**	**332(54,97%)**	**228(46,53%)**	**238(48,77%)**	85(28,33%)
Nouvelle Aquitaine	109(31,87%)	**255(58,22%)**	89(19,26%)	120(19,87%)	205(41,84%)	166(34,02%)	26(8,67%)
Ile de France	**134(39,18%)**	59(13,47%)	178(38,53%)	152(25,16%)	57(11,63%)	84(17,21%)	**189(63,00%)**
Age moyen(E-T)	16,14(1,09)	16,32(1,09)	14,46(1,45)	16,35(1,20)	16,74(1,16)	14,18(0,80)	15,11(1,20)
Age en catégorie							
entre 13 et 15 ans	11(3,22)	1(0,23%)	**280(60,61%)**	30(4,97%)	0(0,00%)	**330(67,83%)**	84(28,00%)
Entre 15 et 18 ans	**296(86,55%)**	**376(85,84%)**	165(35,71%)	**479(79,30%)**	**372(75,92%)**	157(32,17%)	**209(69,67%)**
Plus de 18ans	35(10,23%)	61(13,93%)	17(3,68%)	95(15,73%)	118(24,08%)	1(0,20%)	7(2,33%)
Type d'établissement							
Collège	21(6,23%)	1(0,23%)	**403(88,77%)**	70(11,71%)	2(0,41%)	**472(98,13%)**	**155(52,54%)**
Lycée général et technologique	1(0,30%)	1(0,23%)	6(1,32%)	**521(87,12%)**	10(2,07%)	9(1,87%)	136(46,10)
Lycée professionnel et agricole	**315(93,47%)**	**433(99,54%)**	45(9,91%)	7(1,17%)	**472(97,52%)**	0(0,00%)	4(1,36%)

Légende : **En gras**, type d'établissement le plus représenté dans chaque cluster ; * Cluster identifié à haut risque de symptomatologie dépressive

Tableau 14: Pourcentage d'adolescents à forte symptomatologie dépressive dans chaque cluster, par genre (clusters classés par ordre croissant de proportion d'adolescents avec une ADRS≥6) et résultats de la régression logistique (Odds Ratio symptomatologie dépression~assignation à un cluster)

Garçons			Filles		
CLUSTERS	**% Forte symptomatologie dépressive***	OR[IC95%]	CLUSTERS	% Forte symptomatologie dépressive*	OR[IC95%]
1	1,75	0,29[0,15;0,48]	1	2,03	0,12[0,06;0,2]
2	2,93	0,49[0,29;0,77]	2	8,28	0,52[0,34;0,76]
3	3,18	0,53[0,29;0,88]	3	9,41	0,60[0,45;0,79]
4	3,43	0,57[0,37;0,85]	4	10,37	0,67[0,47;0,93]
5	3,86	0,65[0,39;1,01]	5	11,89	0,78[0,59;1,02]
6	4,49	0,76[0,52;1,08]	6	11,97	0,79[0,54;1,11]
7	4,66	0,79[0,52;1,15]	7	12,33	0,81[0,57;1,12]
8	4,68	0,79[0,47;1,24]	8	12,39	0,82[0,59;1,11]
9	5,02	0,85[0,55;1,26]	9	12,47	0,82[0,58;1,14]
10	**9,31**	**1,66[1,20;2,24]**	10	12,54	0,83[0,59;1,14]
11	**12,25**	**2,25[1,74;2,89]**	11	12,62	0,84[0,59;1,15]
12	**14,49**	**2,74[2,10;3,53]**	12	13,64	0,91[0,64;1,27]
13	**14,96**	**2,84[2,18;3,65]**	13	15,46	1,06[0,78;1,41]
14	**16,00**	**3,07[2,25;4,20]**	14	15,74	1,08[0,85;1,36]
			15	19,52	1,40[1,07;1,82]
			16	21,23	1,56[1,22;1,98]
			17	**29,29**	**2,40[1,99;2,87]**
			18	**29,52**	**2,42[1,94;3,01]**
			19	**31,87**	**2,71[2,19;3,33]**
			20	**44,77**	**4,69[3,69;5,96]**

Légende : * Pourcentage d'adolescent(e)s avec une forte symptomatologie dépressive (score ADRS≥6)
En gras, sont représentés les clusters à « haut risque » tel que défini paragraphe 5.1.2.1 page 100

5.2.2. Description des variables explicatives les plus importantes dans la modélisation

L'analyse par « régression sur profil » a inclus 93 variables explicatives de la symptomatologie dépressive présentes dans l'enquête « Processus d'adolescence », et a permis d'identifier les variables les plus importantes dans le partitionnement.

Les poids de sélection latents pour chaque variable explicative incluse dans l'analyse par ordre croissant en fonction du genre sont présentés en Annexe 4. Parmi les 93 variables inclues, 39 variables différentes ont été identifiées avec un poids de sélection latent supérieur à 0,95 (27 variables en communs et 12 variables différentes : 4 majeures uniquement chez les garçons et 8 majeures uniquement chez les filles). Chez les filles, 35 variables explicatives ont été identifiées avec un poids de sélection supérieur ou égal à 0,95 (moyenne(E-T) =0,73(0,28)), dont 27 avaient un poids supérieur à 0,99 (77%). Tandis que chez les garçons, 31 avaient un poids de sélection supérieur à 0,95 (moyenne(E-T) =0,74(0,27)), parmi elles 23 avec un poids supérieur à 0,99 (74%).

Il est important de noter que le type d'établissement scolaire (collège, lycée général et technologique ou lycée professionnel ou agricole) ainsi que l'âge faisaient partie des variables les plus importantes avec un poids de sélection médian supérieur à 0,998. Ces variables explicatives ont donc joué un rôle majeur dans la constitution des clusters. Cela explique pourquoi, la quasi-totalité des clusters présentaient des profils « homogènes » quant au type d'établissement scolaire. Outre ces caractéristiques scolaires, les variables explicatives majeures dans le modèle s'avéraient être principalement des variables de consommation de substances psychoactives (expérimentation ou consommation intensive), et de sexualité, chez les filles comme chez les garçons.

En revanche, les variables évaluant le redoublement ($poids_{filles}$=0,99 ; $poids_{garcons}$=0,87), la pratique régulière d'un sport ($poids_{filles}$=0,99 ; $poids_{garcons}$=0,89), tout comme la pratique d'un sport à risque ($poids_{filles}$=0,99 ; $poids_{garcons}$=0,60), ou bien la participation à des jeux dangereux ($poids_{filles}$=0,97 ; $poids_{garcons}$=0,89), ont été essentielles dans la constitution des clusters féminins ; alors qu'elles ont contribuées plus faiblement au partitionnement masculin. Chez ces derniers, ce sont les variables « parler le plus facilement de l'actualité avec ses parents qu'avec d'autres » ($poids_{garcons}$=0,97 ; $poids_{filles}$=0,62), et boire de l'alcool plutôt à l'école ou au travail » ($poids_{garcons}$=0,98 ; $poids_{filles}$=0,00), qui ont fait partie des variables explicatives majeures.

Il existe quelques spécificités quant aux modes et types de consommation de substances psychoactives avec un poids très important chez les garçons mais quasi nul chez les filles, par exemple l'expérimentation de drogues par injection avec seringues ($poids_{filles}$=0,03 ; $poids_{garcons}$=0,99) et boire plutôt à l'école ou sur le lieu de travail ($poids_{filles}$=0,00 ; $poids_{garcons}$=0,98).

5.2.3. Description des clusters à « haut risque » de symptomatologie dépressive en fonction des variables explicatives les plus importantes

Les tableaux 15 à 20, décrivent la distribution des variables majeures sus mentionnées dans les différents clusters. Pour rappel, les modèles ont été effectués séparément chez les filles et chez les garçons. Pour une plus grande facilité de présentation, les clusters ayant des profils similaires de scolarisation sont présentés dans les mêmes tableaux. Je m'intéresserai tout d'abord aux filles.

5.2.3.1. Modélisation des clusters à « haut risque » de symptomatologie dépressive chez les filles

Quatre clusters à « haut risque » de symptomatologie dépressive ont été modélisés (clusters 17 à 20, Tableau 14) : Un au collège, deux au lycée général et technologique et un au lycée professionnel ou agricole. Je commencerai par la description de la distribution des variables explicatives, qui sont toutes analysées « à type d'établissement scolaire équivalent » (constitution homogène des clusters en fonction des types d'établissement).

A type d'établissement scolaire équivalent, la consommation de substances psychoactives et la participation à des jeux dangereux, étaient systématiquement les plus fréquentes dans les clusters à « haut risque » (clusters 17 à 20, Tableau 14).

Pour exemple, pour le cluster des collégiennes, la prévalence de participation à des jeux dangereux était de 25,63%, (cluster 20, Tableau 15) 8,60% et 9,97% pour les 2 clusters féminins au lycée général et technologique (respectivement cluster 17 et 18, Tableau 16

Tableau 16), et 20,54% dans le cluster lycée professionnel ou agricole (cluster 19,Tableau 17) ; 63,24% et 68,29% pour les 2 clusters de lycée général et technologique (respectivement cluster 17 et 18, Tableau 16) et 83,25% pour le cluster lycée professionnel ou agricole (cluster 19, Tableau 17). Parmi les 15 autres clusters, les prévalences de participation à des jeux dangereux variaient de 0,19% à 6,33 %. Concernant la sexualité, le fait d'avoir déjà eu un rapport sexuel, était plus fréquent dans les clusters à « haut risque », à type d'établissement scolaire équivalent. Pour le cluster de collégiennes, la prévalence s'élevait à 34,78% (cluster 20, Tableau 15). De même, ces clusters présentaient les prévalences les plus élevées pour les autres variables de sexualité (e.g une attirance homosexuelle ou bisexuelle…).

Concernant le sport, le redoublement et le temps de jeux vidéo, les résultats étaient plus contrastés. Pour exemple, les prévalences de redoublement en collège et lycée général et technologique étaient les plus élevées dans les clusters à « haut risque » : 36,10% pour le cluster collège (cluster 20, Tableau 15), 25,25% et 46,17% pour les 2 clusters lycée général et technologique (respectivement cluster 17 et 18, Tableau 16). Mais cette tendance n'était pas observée dans les clusters à « haut risque » de lycée

professionnel ou agricole (57,66% versus une prévalence variant de 46,12% à 60,39% pour les autres clusters ; Tableau 17).

Description détaillée des 4 clusters à « haut risque »

L'objectif de cette partie de la thèse est de décrire le mieux possible les filles à risque de présenter une forte symptomatologie dépressive. Dans la suite de ce paragraphe, j'effectuerai une description détaillée de ces 4 clusters.

Le cluster 19 (N= 411 moyenne d'âge 16,68 ± 1,23) correspondait au cluster lycée professionnel ou agricole à « haut risque ». Parmi les filles de ce cluster 31,87% d'entre elles présentaient une forte symptomatologie dépressive (Tableau 14).

Les filles du cluster 19 étaient celles parmi les 6 clusters de lycée professionnel ou agricole (Tableau 17), qui avaient le plus déclaré consommer des substances psychoactives et participer à des jeux dangereux. Chez ces filles, 65,77% consommaient intensivement du tabac, 42,54% de l'alcool, 40,70% du cannabis et 91,84% avaient une poly-consommation (consommation régulière d'au moins deux des trois substances psychoactives citées précédemment). De plus, 46,52% avaient expérimenté dans leur vie au moins une autre drogue.

De même, en ce qui concernait les comportements liés à la sexualité, les prévalences de ce cluster à « haut risque » étaient systématiquement les plus élevées des 20 clusters. Les filles de ce cluster étaient relativement nombreuses à avoir redoublé (57,66%), mais sans différence avec les filles des autres clusters lycée professionnel ou agricole. Elles étaient relativement nombreuses à avoir déclaré jouer plus de trois heures par jour aux jeux vidéo et ne pas pratiquer de sport régulièrement. Ce profil pratique de jeux vidéo / sport était partagé par d'autres clusters lycée professionnel ou agricole à « bas risque » (i.e le cluster 11).

Les **clusters 17** (N=594 moyenne d'âge 16,13 ans ± 1,11) et **18** (N= 393 moyenne d'âge 16,32 ans ± 1,31) correspondaient aux deux clusters à « haut risque » regroupant des filles scolarisées en lycée général et technologique. Parmi les filles de ces clusters, 29,29% du cluster 17 et 29,52% du cluster 18, présentaient une forte symptomatologie dépressive (Tableau 14).

Les filles de ces clusters à « haut risque » (clusters 17 et 18) sont celles, qui ont plus fréquemment déclaré consommer des substances psychoactives et participer à des jeux dangereux parmi les 8 clusters de lycée général et technologique (Tableau 16). Entre 22,05% et 38,17% d'entre elles ont déclaré consommé intensivement du tabac, de l'alcool ou du cannabis. Les filles du cluster 17 ont plus fréquemment déclaré consommer des substances que celles du cluster 18 : respectivement 88,47% et 78,67% déclaraient une poly consommation. De plus, 32,19% des filles du cluster 17 ont expérimenté

au moins une autre drogue, 28,76% dans le cluster 18. Les deux tiers des filles de ces clusters ont déjà eu des rapports sexuels (63,24% dans le cluster 17 et 68,29% dans le cluster 18 *versus* prévalence allant de 6,33% à 38,83% dans les autres clusters lycée général et technologique). Notons que parmi tous les clusters de filles scolarisées en lycée général et technologique, celles du cluster 18 ont présenté le taux d'IVG le plus élevé (8,40% d'IVG dans le cluster 18 contre 1,76% dans le cluster 17 et entre 0% et 1,71% dans les autres clusters). Enfin, parmi tous les clusters lycée général et technologique, le cluster 18 est caractérisé par une prévalence de redoublement extrêmement forte (46,17%), et un faible taux de filles faisant du sport (32,06%), *quasi* aucune ne déclaraient faire du sport en compétition (0,25%). Par contre, le cluster 17 avait un niveau de redoublement et un comportement sportif qui se positionnait dans l'étendue des autres clusters de lycée général et technologique (25,25% ont déclaré avoir déjà redoublé et 98,82% ont pratiqué un sport régulièrement). Par rapport aux autres clusters lycées général et technologique, les filles des clusters 17 et 18, n'étaient pas celles qui avaient déclaré jouer le plus aux jeux vidéo durant les vacances scolaires (8,82% et 25,88% respectivement).

Le cluster 20 (N= 277 moyenne d'âge 14,1 ans ± 1,05), correspondait au cluster à « haut risque » au Collège (Tableau 14). Parmi les filles de ce cluster, 44,77% d'entre elles présentaient une forte symptomatologie dépressive.

Les filles de ce cluster (Tableau 15), comme toutes celles des autres clusters à « haut risque », ont déclaré plus fréquemment consommer des substances psychoactives et participer à des jeux dangereux. Il faut noter que 46,15% des filles de ce cluster déclaraient une poly consommation, et 32,12% une expérimentation d'au moins une autre drogue. Le cluster 20 avait la prévalence la plus élevée de redoublement parmi les clusters collège (36,10 % *versus* taux entre 3,33% et 25,67% pour les autres clusters collège). Enfin, ce cluster se caractérisait aussi par un profil sportif avec la quasi-totalité des filles qui déclaraient pratiquer régulièrement du sport (93,41%), dont 54,95% pratiquaient un sport en compétition et 57,52% un sport à risque. Elles étaient en moyenne plus nombreuses à avoir déclaré jouer plus de trois heures par jour aux jeux vidéo durant les vacances par rapport aux autres clusters collège (66,67%), mais tout aussi nombreuses que les filles du cluster 5 qui lui est à « bas risque » de symptomatologie dépressive.

Tableau 15: Description des clusters contenant majoritairement des filles scolarisées au collège en fonction des variables les plus importantes

	Cluster 1	**Cluster 5**	**Cluster 7**	**Cluster 15**	**Cluster 16**	**Cluster 20***
	N=541	N=471	N=300	N=333	N=391	N=277
Forte Symptomatologie depressive(ADRS≥6)	2,03%	11,89%	12,33%	19,52%	21,23%	**44,77%**
Avez-vous déjà redoublé ? (oui)	3,33%	20,81%	25,67%	6,01%	4,86%	**36,10%**
Consommation intensive de tabac durant les 30 derniers jours (≥10 cigarettes/jours)	0,00%	0,00%	0,33%	1,20%	6,14%	**10,14%**
Consommation intensive d'alcool Durant les 30 derniers jours (≥10 verres/mois)	0,19%	0,00%	0,34%	2,86%	8,13%	**9,89%**
Consommation intensive de cannabis Durant les 30 derniers jours (≥10 joints/mois)	0,00%	0,00%	0,00%	1,22%	**5,90%**	**5,17%**
Poly consommation (au moins 2 consommations régulières)	0,00%	0,00%	0,00%	14,70%	12,77%	**46,15%**
Intensité ivresse lors de la dernière fois que vous avez bu (≥5)	0,00%	1,12%	0,72%	7,69%	9,04%	**30,29%**
Quand vous buvez de l'alcool c'est plutôt avec vos amis ou à une soirée ? (oui)	2,77%	13,59%	4,67%	64,26%	30,18%	**74,37%**
Quand vous buvez de l'alcool c'est plutôt seul ? (oui)	0,00%	1,70%	1,33%	3,90%	5,12%	**13,72%**
Quand vous buvez de l'alcool c'est plutôt à l'école, sur votre lieu de stage ou au travail ? (oui) ǂ	0,00%	0,00%	0,33%	0,30%	1,79%	1,44%
Au cours de votre vie, avez-vous consommé des champignons hallucinogènes? (oui)	0,00%	0,22%	0,00%	0,00%	6,15%	7,97%
Au cours de votre vie, avez-vous consommé de l'alcool avec du cannabis ? (oui)	0,00%	0,00%	0,00%	3,95%	9,25%	**25,63%**
Au cours de votre vie, avez-vous consommé de l'alcool avec des médicaments ? (oui)	0,00%	0,00%	0,00%	0,61%	6,15%	**10,55%**
Au cours de votre vie, avez-vous consommé au moins une drogue ? (oui)	0,38%	1,74%	0,34%	9,42%	10,62%	**32,12%**
Avez-vous eu vos règles ? (oui)	66,05%	84,15%	92,23%	90,39%	80,26%	**95,67%**
Participez-vous à des jeux dangereux ? (oui)	0,19%	4,29%	3,01%	6,33%	7,44%	**25,63%**
Hétérosexuelle	91,42%	**95,51%**	82,65%	**96,40%**	92,29%	86,50%
LGB	1,49%	1,50%	2,04%	3,30%	7,46%	**13,50%**
Avez-vous déjà eu des rapports sexuels ? (oui)	0,56%	2,13%	1,68%	7,27%	8,72%	**34,78%**
1er Rapport sexuel précoce dans la relation ? (oui)	0,00%	0,64%	0,67%	0,30%	**2,83%**	2,19%
1er rapport sexuel non protégé ? (oui)	0,18%	0,21%	1,00%	1,80%	2,56%	**7,22%**
IVG (oui)	0,00%	0,00%	0,00%	0,34%	**2,88%**	**3,46%**
Pratiquez-vous régulièrement un sport ?						
Loisir	**60,75%**	55,22%	25,68%	**59,27%**	37,89%	38,46%
Compétition	32,90%	31,74%	3,42%	19,45%	**60,05%**	54,95%
Pratiquez-vous un sport à risque?	22,18%	37,28%	0,36%	1,56%	**62,99%**	57,52%
Connaitre ses limites pratique d'un sport à risque	27,57%	37,14%	2,79%	2,17%	**59,64%**	49,81%
Parlez–vous facilement avec vos parents de l'actualité?ǂ	67,10%	50,78%	26,77%	47,26%	38,28%	32,77%
Temps de jeux vidéo en semaine						
Entre 1H et 3H par jour	5,93%	47,79	26,34%	9,43%	4,79%	41,74%
Plus de 3h par jour	0,00%	**15,93%**	3,57%	0,00%	2,24%	**13,91%**
Temps de jeux vidéo en week-end						
Entre 1H et 3H par jour	36,07%	57,11%	39,48%	37,45%	21,74%	54,70%
Plus de 3h par jour	0,00%	**41,93%**	17,17%	0,80%	2,80%	38,03%
Temps de jeux vidéo pendant les vacances						
Entre 1H et 3H par jour	46,76%	28,92%	40,33%	46,54%	35,20%	27,23%
Plus de 3h par jour	3,94%	**68,19%**	25,21%	6,15%	7,17%	**67,66%**

Légende : **en gras,** le pourcentage le plus élevé pour chaque variable explicative. * Cluster à haut risque, tel que défini paragraphe 5.1.2.1 page 100 ; ǂ Variable majeure uniquement chez les garçons.

Tableau 16: Description des clusters contenant majoritairement des filles scolarisées en lycée général ou technologique en fonction des variables les plus importantes

	Cluster 2 N=302	**Cluster 3** N=563	**Cluster 4** N=328	**Cluster10** N=311	**Cluster 12** N=264	**Cluster 13** N=317	**Cluster 17*** N=594	**Cluster 18*** N=393
Forte symptomatologie depressive(ADRS≥6)	8,28%	9,41%	10,37%	12,54%	13,64%	15,46%	29,29%	**29,52%**
Avez-vous déjà redoublé ? (oui)	18,27%	12,63%	18,96%	20,58%	29,92%	23,81%	25,25%	**46,17%**
Consommation intensive de tabac durant les 30 derniers jours (≥10 cigarettes/jours)	0,00%	1,95%	0,91%	1,93%	1,52%	0,63%	22,05%	**38,17%**
Consommation intensive d'alcool durant les 30 derniers jours (≥10 verres/mois)	0,00%	6,16%	3,79%	10,39%	3,57%	0,00%	**30,31%**	24,74%
Consommation intensive de cannabis Durant les 30 derniers jours (≥10 joints/mois)	0,00%	0,89%	0,61%	0,97%	0,76%	0,00%	**31,01%**	29,84%
Poly consommation (au moins 2 consommations régulières)	0,00%	16,30%	14,20%	32,03%	5,58	0,64%	**88,47%**	78,67%
Intensité ivresse lors de la dernière fois que vous avez bu (≥5)	0,00%	31,17%	25,00%	31,39%	8,56%	0,66%	**79,70%**	64,71%
Quand vous buvez de l'alcool c'est plutôt avec vos amis ou à une soirée ? (oui)	12,91%	92,72%	79,88%	94,21%	55,68%	7,26%	**99,83%**	96,18%
Quand vous buvez de l'alcool c'est plutôt seul ? (oui)	1,32%	0,89%	1,22%	0,96%	1,52%	0,32%	2,02%	**3,56%**
Quand vous buvez de l'alcool c'est plutôt à l'école, sur votre lieu de stage ou au travail ? (oui)ǂ	0,00%	0,89%	0,92%	1,61%	0,76%	0,00%	2,53%	**4,07%**
Au cours de votre vie, avez-vous consommé des champignons hallucinogènes ? (oui)	0,00%	0,18%	0,61%	0,00%	0,00%	0,00%	**12,37%**	8,57%
Au cours de votre vie, avez-vous consommé de l'alcool avec du cannabis ? (oui)	0,00%	9,45%	7,65%	20,26%	0,38%	0,32%	**76,36%**	61,72%
Au cours de votre vie, avez-vous consommé de l'alcool avec des médicaments ? (oui)	0,33%	0,71%	1,83%	0,00%	0,00%	0,00%	**11,05%**	6,94%
Au cours de votre vie, avez-vous consommé au moins une drogue ? (oui)	1,35%	2,50%	2,76%	5,18%	2,32%	0,00%	**32,19%**	28,76%
Avez-vous eu vos règles ? (oui)	97,67%	98,93%	**99,70%**	97,43%	99,62%	98,74%	99,66%	99,74%
Participez-vous à des jeux dangereux ? (oui)	0,33%	0,36%	2,44%	1,29%	1,15%	0,95%	8,60%	**9,97%**
Attirance sexuelle								
Hétérosexuelle	95,36%	95,37%	93,58%	94,52%	91,63%	**95,57%**	89,73%	90,08%
LGB	3,31%	4,63%	6,12%	5,48%	6,46%	2,85%	**10,27%**	9,67%
Avez-vous déjà eu des rapports sexuels ? (oui)	6,33%	37,21%	28,75%	38,83%	21,46%	9,46%	63,24%	**68,29%**
1er Rapport sexuel précoce dans la relation ? (oui)	0,33%	0,00%	0,92%	1,29%	1,15%	0,32%	4,94%	**10,08%**
1er rapport sexuel non protégé ? (oui)	0,00%	0,71%	1,52%	1,29%	1,52%	0,32%	6,23%	**13,74%**
IVG (oui)	0,00%	1,34%	0,66%	1,71%	0,41%	0,35%	1,76%	**8,40%**
Pratiquez-vous régulièrement un sport ?								
Loisir	61,92%	**69,64%**	65,55%	42,12%	27,48%	46,82%	61,55%	31,81%
Compétition	37,75%	20,71%	27,44%	**57,88%**	0,76%	8,28%	37,27%	0,25%
Pratiquez-vous un sport à risque?	51,01%	1,28%	38,80%	**99,04%**	1,63%	1,99%	39,12%	0,82%
Connaitre ses limites pratique d'un sport à risque	56,23%	5,61%	37,12%	**89,39%**	1,94%	3,25%	35,48%	2,05%
Parlez–vous facilement avec vos parents de l'actualité?ǂ	71,70%	**92,72%**	79,88%	65,14%	66,38%	42,43%	52,91%	41,31%
Temps de jeux vidéo en semaine								
Entre 1H et 3H par jour	6,81%	0,00%	**33,76%**	1,24%	12,96%	3,13%	3,15%	10,42%
Plus de 3h par jour	0,00%	0,00%	**5,79%**	0,00%	2,78%	0,00%	0,45%	1,30%
Temps de jeux vidéo en week-end								
Entre 1H et 3H par jour	32,66%	8,00%	**68,24%**	14,69%	54,63%	33,05%	23,78%	31,01%
Plus de 3h par jour	1,21%	0,00%	**26,10%**	0,41%	11,45%	0,00%	1,33%	11,39%
Temps de jeux vidéo pendant les vacances								
Entre 1H et 3H par jour	40,16%	26,94%	**39,50%**	34,54%	42,86%	45,27%	36,34%	29,07%
Plus de 3h par jour	6,83%	0,68%	**57,99%**	2,01%	34,20%	11,93%	8,82%	25,88%

Légende : **en gras**, le pourcentage le plus élevé pour chaque variable explicative. * Cluster à haut risque, tel que défini paragraphe 5.1.2.1 page 100 ; ǂ Variable majeure uniquement chez les garçons

Tableau 17: Description des clusters contenant des filles majoritairement scolarisées en «lycée professionnel et agricole» en fonction des variables les plus importantes

	Cluster 6	Cluster 8	Cluster 9	Cluster 11	Cluster 14	Cluster 19*
	N=259	N=339	N=289	N=309	N=521	N=411
Forte symptomatologie depressive(ADRS≥6)	11,97%	12,39%	12,47%	12,62%	15,74%	31,87%
Avez-vous déjà redoublé ? (oui)	46,12%	49,85%	53,63%	60,39%	**62,57%**	57,66%
Consommation intensive de tabac durant les 30 derniers jours (≥10 cigarettes/jours)	3,09%	27,73%	13,15%	1,95%	13,87%	**65,77%**
Consommation intensive d'alcool Durant les 30 derniers jours (≥10 verres/mois)	0,40%	25,15%	17,83%	1,65%	13,47%	**42,54%**
Consommation intensive de cannabis Durant les 30 derniers jours (≥10 joints/mois)	0,39%	7,42%	3,50%	0,00%	0,78%	**40,70%**
Poly consommation (au moins 2 consommations régulières)	3,97%	54,79%	38,16%	1,67%	29,80%	**91,84%**
Intensité ivresse lors de la dernière fois que vous avez bu (≥5)	2,37%	47,92%	31,69%	0,34%	25,64%	**71,99%**
Quand vous buvez de l'alcool c'est plutôt avec vos amis ou à une soirée ? (oui)	14,67%	**98,53%**	83,74%	17,48%	81,57%	93,92%
Quand vous buvez de l'alcool c'est plutôt seul ? (oui)	0,00%	4,42%	2,08%	0,00%	0,78%	**4,87%**
Quand vous buvez de l'alcool c'est plutôt à l'école, sur votre lieu de stage ou au travail ?ǂ (oui)	14,67%	98,53%	83,74%	17,48%	81,57%	**93,92%**
Au cours de votre vie, avez-vous consommé des champignons hallucinogènes? (oui)	0,00%	4,18%	2,79%	0,00%	0,00%	**19,26%**
Au cours de votre vie, avez-vous consommé de l'alcool avec du cannabis ? (oui)	0,78%	35,61%	10,07%	0,33%	8,90%	**76,17%**
Au cours de votre vie, avez-vous consommé de l'alcool avec des médicaments ? (oui)	0,39%	2,10%	3,48%	0,00%	0,58%	**15,69%**
Au cours de votre vie, avez-vous consommé au moins une drogue ? (oui)	1,57%	14,07%	7,72%	1,66%	3,53%	**46,52%**
Avez-vous eu vos règles ? (oui)	98,84%	100,00%	98,96%	98,05%	100,00%	99,76%
Participez-vous à des jeux dangereux ? (oui)	2,70%	4,13%	4,84%	2,28%	1,15%	**20,54%**
Attirance sexuelle						
Hétérosexuelle	89,53%	94,69%	89,97%	93,49%	95,58%	**85,33%**
LGB	5,81%	5,31%	9,69%	3,91%	4,04%	**14,67%**
Avez-vous déjà eu des rapports sexuels ? (oui)	25,87%	73,96%	67,82%	24,18%	56,81%	**83,25%**
1er Rapport sexuel précoce dans la relation ? (oui)	0,00%	3,88%	3,82%	0,33%	1,55%	**8,98%**
1er rapport sexuel non protégé ? (oui)	3,86%	3,24%	9,00%	2,27%	4,22%	**18,00%**
IVG (oui)	1,22%	7,14%	2,83%	3,52%	2,02%	**10,47%**
Pratiquez-vous régulièrement un sport ?						
Loisir	**64,06%**	51,92%	54,90%	22,26%	28,21%	41,63%
Compétition	29,30%	**47,79%**	38,11%	6,64%	0,38%	9,61%
Pratiquez-vous un sport à risque?	44,84%	**58,81%**	49,29%	3,53%	0,82%	15,98%
Connaitre ses limites pratique d'un sport à risque	46,99%	**55,65%**	47,69%	5,05%	0,58%	12,69%
Parlez–vous facilement avec vos parents de l'actualité?ǂ	53,25%	47,24%	**65,46%**	55,19%	47,62%	41,62%
Temps de jeux vidéo en semaine						
Entre 1H et 3H par jour	2,49%	2,08%	**35,34%**	36,25%	5,00%	20,43%
Plus de 3h par jour	1,00%	0,00%	**7,14%**	5,24%	0,26%	5,79%
Temps de jeux vidéo en week-end						
Entre 1H et 3H par jour	21,56%	11,55%	**68,86%**	56,97%	25,69%	41,16%
Plus de 3h par jour	3,67%	0,00%	**30,40%**	26,23%	1,75%	22,61%
Temps de jeux vidéo pendant les vacances						
Entre 1H et 3H par jour	33,64%	36,50%	**42,12%**	42,51%	30,22%	33,62%
Plus de 3h par jour	7,83%	0,00%	**56,04%**	40,08%	10,32%	37,64%

Légende : **en gras**, le pourcentage le plus élevé pour chaque variable explicative. * Cluster à haut risque, tel que défini paragraphe 5.1.2.1page 100 ; ǂ Variable majeure uniquement chez les garçons

5.2.3.2. Modélisation des clusters chez les garçons

Je m'intéresserai tout d'abord à la distribution des variables explicatives dans les 5 clusters à « haut risque » de symptomatologie dépressive identifiés chez les garçons (clusters 10 à 14, Tableau 14).

Chez les garçons, les variables explicatives de pratique régulière d'un sport, de redoublement et de participation à des jeux dangereux, n'ont pas été identifiées comme variables majeures du partitionnement (contrairement aux filles).

Les taux de garçons ayant déjà eu des rapports sexuels étaient systématiquement les plus élevés dans les clusters à « haut risque », à type d'établissement équivalent. Pour les clusters de collégiens, le taux variait de 19,09% à 59,47% (clusters 10 à 13, Tableau 18), 77,28% pour le cluster de lycée général et technologique (Tableau 19) et 87,81% pour le cluster de lycée professionnel ou agricole (Tableau 20). De même, ces clusters présentaient, les taux les plus élevés sur les autres variables de sexualité.

En ce qui concerne les variables de consommation de substances psychoactives, les clusters de lycées général et technologique et lycée professionnel ou agricole, présentaient les prévalences les plus élevées : respectivement 30,30% (cluster 11, Tableau 19) et 53,18% (cluster 12, Tableau 20) déclaraient avoir fumé intensivement durant le mois ayant précédé l'enquête.

En revanche, pour les clusters de collégiens (Tableau 18), les résultats sont plus complexes à interpréter : seuls les clusters 10 et 13 étaient caractérisés par des prévalences plus élevées de consommation intensive de tabac, alcool ou cannabis dans le mois précédant l'enquête et d'expérimentation des autres drogues comparativement aux autres clusters de collège dont le cluster 14 (cluster à « haut risque »).

Par ailleurs, les clusters à « haut risque », tout type d'établissement confondu, se caractérisaient par les taux les plus élevés de sujets déclarant avoir déjà bu à l'école ou sur leur lieu de stage ou au travail, avec des prévalences allant de 7,69% et 12,24% pour les clusters de lycée général et technologique (clusters 11 ; Tableau 19) et lycée professionnel ou agricole (cluster 12,Tableau 20) et de 4,30% et 6,71% pour les 2 clusters de collégiens qui consommaient (clusters 13 et 10 Tableau 18).

Description détaillée des 5 clusters à « haut risque ».

Dans la suite de ce paragraphe, j'effectuerai une description détaillée des 5 clusters à « haut risque », afin de dresser le profil des garçons présents dans ces clusters.

Le cluster 12 (N= 490 moyenne d'âge 16,7 ± 1,2) correspondait au cluster lycée professionnel ou agricole à « haut risque », 14,49% des garçons de ce cluster présentaient une forte symptomatologie dépressive (Tableau 14). Ils déclaraient plus fréquemment consommer des substances psychoactives (Tableau 20). La majorité des garçons de ce cluster consommaient intensivement au moment de l'enquête de l'alcool, du tabac ou du cannabis et 93,62% déclaraient une poly-consommation (consommation régulière d'au moins deux des trois drogues citées précédemment). Chez ces garçons, 42,71% avaient expérimenté dans leur vie au moins une autre drogue. De même, les taux de comportements liés à la sexualité, sont systématiquement les plus élevés des 14 clusters (pour exemple, 87,81% des garçons de ce cluster, ont déclaré avoir déjà eu des rapports sexuels, *versus* un taux variant de 35,10% à 65,65% pour les autres clusters de ce type d'établissement). Ils étaient nombreux à avoir déclaré jouer plus de trois heures par jour aux jeux vidéo comparés aux autres clusters lycée professionnel ou agricole, mais moins nombreux que les garçons du cluster 9 à « bas risque » (Tableau 20). En revanche, ils étaient moins nombreux à avoir déclaré discuter le plus facilement de l'actualité avec leurs parents qu'avec d'autres (42,55% dans le cluster 12 *versus* une prévalence supérieure à 50% dans les autres clusters).

Le cluster 11 (N= 604 moyenne d'âge 16,4 ± 1,2), correspondait au seul cluster lycée général et technologique à « haut risque ». Parmi les garçons de ce cluster, 12,25% présentaient une forte symptomatologie dépressive (Tableau 14). Ils déclaraient consommer des substances psychoactives : 89,42% déclaraient une poly-consommation, et 27,40% une expérimentation d'au moins une autre drogue (Tableau 19). Parmi les garçons de ce cluster, 77,28% ont déclaré avoir déjà eu un rapport sexuel et 5,63%, un premier rapport sexuel non protégé. Ils étaient peu nombreux à déclarer jouer aux jeux vidéo durant plus de trois heures pendant la semaine (2,26%), bien moins que le cluster 6 (cluster à « bas risque » présentant les prévalences les plus élevées de temps de jeux vidéo pendant plus de 3h par jour).

Les clusters 10 (N= 462 moyenne d'âge 14,5 ans ±1,2) **13** (N=488 moyenne d'âge 14,2 ans ±0,8) et **14** (N= 300 moyenne d'âge 15,1 ans ±1,2) correspondaient aux trois clusters collège à « haut risque ». Le cluster 14 était un cluster mixte contenant à la fois des collégiens et des lycéens (52,46% de collégiens). Parmi les garçons de ces clusters, 9,31%, 14,96% et 16,00% (respectivement cluster 10 ,13 et 14) présentaient une forte symptomatologie dépressive (Tableau 14).

Parmi les clusters collège (Tableau 18), les garçons des clusters 10 et 13 présentaient les prévalences de consommation de substances les plus élevées : 57,93% des garçons du cluster 13 déclaraient une poly consommation. Ils consommaient plus fréquemment que ceux du cluster 10 (18,12% déclaraient une poly-consommation). De plus, 24,79% des garçons du cluster 13 ont expérimenté au moins une autre drogue ainsi que 16,85% dans le cluster 10. Toutefois, il faut noter que les garçons du cluster 10 déclaraient pour l'expérimentation d'amphétamines et de champignons hallucinogènes les prévalences les plus élevées. Bien que le cluster 14 fasse partie du trio des clusters collège les plus à risque en ce qui concerne les consommations de substances, ce cluster a un profil très similaire aux clusters à « bas risque » (cluster contenant significativement moins des garçons avec une forte symptomatologie dépressive). Les garçons du cluster 10 ont davantage déclaré avoir une attirance homosexuelle ou bisexuelle par rapport à ceux des autres clusters. Ils ont notamment plus déclaré avoir déjà eu des rapports (19,09% pour le cluster 10, 59,47% pour le cluster 13 et 41,14% pour le cluster 14 *versus* un taux variant de 3,17% et 17,38% dans les autres clusters collège). Cependant, les garçons du cluster 14 ont plus déclaré une sexualité à risque que les autres garçons. Leur prévalence est plus élevée que celles du cluster 10 et 13, d'avoir eu leur 1er rapport avec quelqu'un qu'il venait de rencontrer (10,47% dans le cluster 14, 8,70% et 8,87% dans les clusters 10 et 13 et <2% dans les autres clusters) ; d'avoir eu un 1er rapport non protégé (11,00% dans le cluster 14, 6,93% et 7,79% dans les clusters 10 et 13 et <2% dans les autres clusters) ; et leurs petites amies d'avoir plus eu recours à une IVG (4,42% dans le cluster 14, 2,90% et 3,54% dans les clusters 10 et 13 et <1,2% dans les autres clusters).

Tableau 18: Description des clusters regroupant majoritairement des garçons scolarisés au collège selon les variables les plus importantes

	Cluster 2	Cluster 3	Cluster 7	Cluster 10*	Cluster 13*	Cluster 14*
	N=512	N=377	N=515	N=462	N=488	N=300
Forte symptomatologie depressive (ADRS≥6)	2,93%	3,18%	4,66%	9,31%	14,96%	16,00%
Avez-vous déjà redoublé ? (oui) ǂ	8,20%	7,47%	10,87%	31,89%	**48,98%**	48,32%
Consommation intensive de tabac durant les 30 derniers jours (≥10 cigarettes/jours)	0,00%	0,00%	0,19%	12,17%	**18,93%**	3,01%
Consommation intensive d'alcool durant les 30 derniers jours (≥10 verres/mois)	0,20%	3,39%	1,44%	16,21%	**32,40%**	0,35%
Consommation intensive de cannabis durant les 30 derniers jours (≥10 joints/mois)	0,00%	1,07%	0,00%	**12,04%**	10,32%	2,37%
Poly-consommation (au moins 2 consommations régulières)	0,40%	9,38%	1,24%	18,12%	**57,93%**	2,12%
Intensité ivresse lors de la dernière fois que vous avez bu (≥5)	0,61%	9,26%	2,42%	16,37%	**48,54%**	0,36%
Quand vous buvez de l'alcool c'est plutôt seul ? (oui) ǂ	1,17%	6,10%	2,14%	8,87%	**12,09%**	1,67%
Quand vous buvez de l'alcool c'est plutôt en avec des copains ou en soirée ? (oui)	6,05%	55,44%	20,58%	25,32%	**85,04%**	4,67%
Quand vous buvez de l'alcool c'est plutôt à l'école, sur votre lieu de stage ou au travail ? (oui)	0,59%	1,33%	0,19%	**6,71%**	4,30%	0,00%
Au cours de votre vie, avez-vous consommé des amphétamines ? (oui)	0,20%	0,54%	1,18%	**9,98%**	1,48%	2,40%
Au cours de votre vie, avez-vous consommé des champignons hallucinogènes ? (oui)	0,39%	0,54%	0,20%	**13,44%**	8,47%	0,68%
Au cours de votre vie avez-vous consommé au moins une autre drogue ? (oui)	1,20%	5,41%	2,77%	16,85%	**24,79%**	5,19%
Au cours de votre vie, avez-vous consommé de l'alcool avec du cannabis ? (oui)	0,59%	5,08%	0,39%	15,16%	**33,89%**	1,02%
Attirance sexuelle						
Hétérosexuelle	92,67%	98,93%	**99,22%**	87,36%	95,69%	97,00%
LGB	0,40%	1,07%	0,59%	**4,36%**	3,90%	2,67%
Avez-vous déjà eu des rapports sexuels ? (oui)	3,17%	17,38%	10,37%	19,09%	**59,47%**	41,14%
1er Rapport sexuel précoce dans la relation ? (oui)	0,79%	1,87%	0,39%	8,70%	8,87%	**10,47%**
1er rapport sexuel non protégé ?(oui)	0,59%	1,86%	0,78%	6,93%	7,79%	**11,00%**
IVG (oui)	0,00%	1,19%	0,00%	2,90%	3,54%	**4,42%**
Pratiquez-vous régulièrement un sport ? ǂ						
loisir	**41,80%**	21,10%	37,92%	36,75%	47,30%	38,23%
compétition	**48,63%**	76,99%	60,28%	26,06%	36,72%	55,63%
Pratiquez-vous un sport à risque? ǂ	**29,01%**	51,14%	38,82%	18,48%	51,50%	38,75%
Connaitre ses limites pratique d'un sport à risque ǂ	**29,98%**	45,92%	39,10%	17,04%	40,30%	40,83%
Parlez–vous facilement avec vos parents de l'actualité?	**67,02%**	60,27%	66,83%	43,85%	48,03%	29,55%
Temps de jeux vidéo en semaine						
Entre 1H et 3H par jour	22,56%	21,41%	**52,52%**	49,49%	51,08%	43,43%
Plus de 3h par jour	0,75%	1,92%	**30,73%**	30,71%	29,26	24,30%
Temps de jeux vidéo en week-end						
Entre 1H et 3H par jour	70,55%	64,00%	**22,30%**	25,68%	32,95%	36,65%
Plus de 3h par jour	1,43%	4,31%	**77,48%**	70,86%	69,12%	56,97%
Temps de jeux vidéo pendant les vacances						
Entre 1H et 3H par jour	56,80%	59,77%	**5,66%**	13,90%	13,95%	16, 47%
Plus de 3h par jour	20,05%	24,00%	**92,31%**	82,38%	83,45%	78,82%

Légende : **en gras**, le pourcentage le plus élevé pour chaque variable explicative. * Cluster à « haut risque », tel que défini paragraphe 5.1.2.1 page 100 ; ǂ Variable majeure uniquement chez les filles

Tableau 19: Description des clusters contenant majoritairement des garçons scolarisés en lycée général ou technologique selon les variables les plus importantes

	Cluster 4 N=612	**Cluster 5** N=414	**Cluster 6** N=623	**Cluster 11*** N=604
Forte symptomatologie depressive (ADRS≥6)	3,43%	3,86%	4,49%	12,25%
Avez-vous déjà redoublé ? (oui) ǂ	24,84%	18,40%	20,26%	**39,93%**
Consommation intensive de tabac durant les 30 derniers jours (≥10 cigarettes/jours)	5,56%	0,48%	3,05%	**30,30%**
Consommation intensive d'alcool durant les 30 derniers jours (≥10 verres/mois)	24,92%	4,28%	18,11%	**47,14%**
Consommation intensive de cannabis durant les 30 derniers jours (≥10 joints/mois)	4,95%	0,24%	3,86%	**40,40%**
Poly-consommation (au moins 2 consommations régulières)	15,22%	1,76%	23,77%	**89,42%**
Intensité ivresse lors de la dernière fois que vous avez bu (≥5)	37,85%	1,74%	38,05%	**77,57%**
Quand vous buvez de l'alcool c'est plutôt seul ? (oui) ǂ	1,63%	1,45%	5,30%	**5,80%**
Quand vous buvez de l'alcool c'est plutôt en avec des copains ou en soirée ? (oui)	88,73%	21,74%	84,59%	**99,17%**
Quand vous buvez de l'alcool c'est plutôt à l'école, sur votre lieu de stage ou au travail ? (oui)	3,43%	0,72%	2,09%	**7,69%**
Au cours de votre vie, avez-vous consommé des amphétamines ? (oui)	1,64%	0,00%	0,64%	**2,72%**
Au cours de votre vie, avez-vous consommé des champignons hallucinogènes ? (oui)	4,59%	0,00%	0,80%	**14,68%**
Au cours de votre vie avez-vous consommé au moins une autre drogue ? (oui)	6,44%	0,49%	3,25%	**27,40%**
Au cours de votre vie, avez-vous consommé de l'alcool avec du cannabis ? (oui)	12,30%	0,49%	19,00%	**84,25%**
Attirance sexuelle				
Hétérosexuelle	**99,02%**	95,63%	95,65%	96,68%
LGB	0,98%	3,40%	**4,03%**	3,32%
Avez-vous déjà eu des rapports sexuels ? (oui)	49,18%	6,07%	36,14%	**77,28%**
1er Rapport sexuel précoce dans la relation ? (oui)	7,45%	0,24%	4,56%	**20,23%**
1er rapport sexuel non protégé ? (oui)	2,45%	0,24%	1,44%	**5,63%**
IVG (oui)	2,22%	0,25%	1,62%	**7,09%**
Pratiquez-vous régulièrement un sport ? ǂ				
loisir	**31,91%**	48,17%	48,71%	44,07%
compétition	**67,11%**	33,50%	41,80%	50,58%
Pratiquez-vous un sport à risque? ǂ	**61,06%**	21,50%	38,82%	51,11%
Connaitre ses limites pratique d'un sport à risque ǂ	**53,83%**	27,43%	34,59%	43,15%
Parlez–vous facilement avec vos parents de l'actualité?	**71,54%**	71,35%	63,47%	51,74%
Temps de jeux vidéo en semaine				
Entre 1H et 3H par jour	**16,17%**	21,49%	53,78%	25,85%
Plus de 3h par jour	2,42%	0,57%	**16,30%**	2,26%
Temps de jeux vidéo en week-end				
Entre 1H et 3H par jour	64,08%	58,81%	**23,54%**	51,66%
Plus de 3h par jour	5,05%	10,03%	**76,46%**	16,79%
Temps de jeux vidéo pendant les vacances				
Entre 1H et 3H par jour	53,57%	48,24%	**5,03%**	41,36%
Plus de 3h par jour	22,68%	27,10%	**93,30%**	34,56%

Légende : **en gras**, le pourcentage le plus élevé pour chaque variable explicative.* Cluster à « haut risque », tel que défini paragraphe 5.1.2.1 page 100 Caractérisation des différents profils de cluster à « haut risque »; ǂ variables majeures uniquement chez les filles

Tableau 20: Description des clusters contenant majoritairement des garçons scolarisés en lycée professionnel ou agricole selon les variables explicatives

	Cluster 1	**Cluster 8**	**Cluster 9**	**Cluster 12***
	N=628	N=342	N=438	N=490
Forte symptomatologie depressive (ADRS≥6)	1,75%	4,68%	5,02%	14,49%
Avez-vous déjà redoublé ? (oui) ǂ	44,59%	53,22%	46,80%	**57,35%**
Consommation intensive de tabac durant les 30 derniers jours (≥10 cigarettes/jours)	21,97%	1,47%	4,12%	**53,18%**
Consommation intensive d'alcool durant les 30 derniers jours (≥10 verres/mois)	47,15%	0,91%	28,64%	**53,01%**
Consommation intensive de cannabis durant les 30 derniers jours (≥10 joints/mois)	9,63%	1,19%	0,00%	**59,33%**
Poly-consommation (au moins 2 consommations régulières)	51,88%	2,15%	19,14%	**93,62%**
Intensité ivresse lors de la dernière fois que vous avez bu (≥5)	59,26%	0,00%	33,33%	**74,27%**
Quand vous buvez de l'alcool c'est plutôt seul ? (oui) ǂ	**1,12%**	1,46%	5,48%	9,59%
Quand vous buvez de l'alcool c'est plutôt en avec des copains ou en soirée ? (oui)	**96,82%**	6,43%	77,40**%**	74,27%
Quand vous buvez de l'alcool c'est plutôt à l'école, sur votre lieu de stage ou au travail ? (oui)	7,48%	0,88%	6,62%	**12,24%**
Au cours de votre vie, avez-vous consommé des amphétamines ? (oui)	0,64%	0,00%	0,69%	**4,41%**
Au cours de votre vie, avez-vous consommé des champignons hallucinogènes ? (oui)	3,22%	0,30%	0,69%	**25,00%**
Au cours de votre vie avez-vous consommé au moins une autre drogue ? (oui)	6,13%	0,90%	3,25%	**42,71%**
Au cours de votre vie, avez-vous consommé de l'alcool avec du cannabis ? (oui)	29,76%	0,60%	6,90%	**83,68%**
Attirance sexuelle				
Hétérosexuelle	99,36%	92,08%	97,48%	94,05%
LGB	0,64%	3,52%	2,52%	**5,54%**
Avez-vous déjà eu des rapports sexuels ? (oui)	65,65%	35,10%	47,36%	**87,81%**
1er Rapport sexuel précoce dans la relation ? (oui)	9,16%	8,85%	3,70%	**22,06%**
1er rapport sexuel non protégé ? (oui)	2,07%	4,39%	2,28%	**10,61%**
IVG (oui)	3,33%	0,72%	3,13%	**12,47%**
Pratiquez-vous régulièrement un sport ? ǂ				
loisir	43,64%	38,30%	44,39%	**47,30%**
compétition	48,63%	45,59%	38,32%	**36,72%**
Pratiquez-vous un sport à risque? ǂ	54,28%	33,33%	36,12%	**44,52%**
Connaitre ses limites pratique d'un sport à risque ǂ	47,55%	35,09%	37,86%	**35,88%**
Parlez–vous facilement avec vos parents de l'actualité?	**64,34%**	53,49%	61,46%	42,55%
Temps de jeux vidéo en semaine				
Entre 1H et 3H par jour	**18,92%**	35,64%	38,73%	36,28%
Plus de 3h par jour	0,39%	17,09%	**23,04%**	22,91%
Temps de jeux vidéo en week-end				
Entre 1H et 3H par jour	47,79%	41,70%	**34,82%**	32,95%
Plus de 3h par jour	6,09%	42,76%	**63,46%**	56,68%
Temps de jeux vidéo pendant les vacances				
Entre 1H et 3H par jour	40,74%	31,45%	**14,57%**	21,11%
Plus de 3h par jour	20,56%	54,77%	**84,20%**	70,77%

Légende : **en gras**, le pourcentage le plus élevé pour chaque variable explicative. * Cluster à « haut risque », tel que défini paragraphe 5.1.2.1 page 100 ; ǂ variables majeures uniquement chez les filles

5.3. Discussion

Dans ce chapitre de thèse, je me suis intéressée à l'identification et à la caractérisation de profils différents d'adolescents à risque de présenter une forte symptomatologie dépressive. Ces profils ont été constitués à partir d'une large gamme de variables explicatives sociodémographiques (âge, statut scolaire, structure familiale, niveau d'éducation du père et de la mère), liées au mode de vie (consommation de substances, loisirs, sport, expérience sexuelle, sommeil), aux relations familiales et amicales et au corps (IMC, alimentation, image corporelle…). Dans ce chapitre, j'ai utilisé une méthode de partitionnement guidée par les données, la « régression sur profil », méthode supervisée permettant d'analyser l'effet conjoint d'un ensemble de variables explicatives sur une variable d'intérêt. L'utilisation de la sélection de variables a permis de classer les variables explicatives par ordre d'importance et de décrire les clusters de façon pertinente en se focalisant sur les variables ayant été majeures dans la structuration des clusters.

Sur le plan méthodologique, l'utilisation de la « régression sur profil » a été un point fort de cette analyse. Il s'agit d'une approche validée en épidémiologie dans des domaines variés, allant de l'épidémiologie génétique (136) à l'épidémiologie clinique en passant par l'épidémiologie environnementale (133,137). Par exemple, elle a permis de dresser les phénotypes de contrôle cardio-ventilatoires associés à la présence d'une apnée obstructive du sommeil durant l'enfance chez les enfants nés prématurément (138), d'explorer les interactions gène-environnement sur le risque de cancer du poumon (136), d'identifier les profils de consommation alimentaire des sous-classes de polyphénols associées au risque de diabète de type 2 (139). L'étude princeps de Molitor et al (2010) décrivait une application de la régression sur profil pour identifier les combinaisons de facteurs familiaux et environnementaux à risque de problème de santé mentale chez des enfants âgés entre 6 et 17 ans (75).

Au total, 34 clusters ont été constitués par la « régression sur profil » : 20 chez les filles dont 4 à « haut risque » et 14 chez les garçons dont 5 à « haut risque ». L'étape de sélection de variables a mis en évidence, 35 variables chez les filles et 31 variables chez les garçons, particulièrement importantes dans la structuration des clusters sur les 93 variables incluent dans le modèle. Dans la suite de ce chapitre, je vais discuter les résultats marquants sur le plan de l'interprétation méthodologique et de l'intérêt de la régression sur profil. Ces résultats seront discutés sur le plan clinique et comparés à la littérature sur les facteurs de risque de la dépression dans le chapitre de discussion générale de thèse de façon à être analysés en regard des résultats du chapitre 4 .

Le type d'établissement scolaire s'est avéré être une variable majeure dans l'élaboration des profils. Par exemple, certains clusters contenaient jusqu'à 100% d'adolescents scolarisés dans un même type d'établissement. Sans stratification a priori, mon analyse a donc regroupé dans les mêmes clusters des adolescents ayant des profils scolaires similaires ; reflétant ainsi le panorama de l'enseignement en France. Ce découpage des données, m'a interrogé sur sa signification et la manière d'interpréter les résultats dans ces clusters.

En effet, la variable « établissement scolaire » contient à la fois des notions d'âge (différence collège/lycée), de niveau socioculturel (général/professionnel/agricole), d'implantation géographique et de proportion filles/garçons. Par exemple, la proportion d'élèves de plus de 18 ans étant bien plus importante en lycée professionnel ou agricole qu'en lycée général et technologique (25,56% *vs* 8,76% chez les filles ; 17,91% *vs* 9,81% chez les garçons). De même, pour la consommation de substances psychoactives. Les adolescents scolarisés en enseignement professionnel ou agricole consommaient plus que ceux scolarisés en enseignement général (140,141). La variable « établissement scolaire » présente donc à la fois un risque de sur-ajustement en cas de corrélation avec des variables proches et une multisémie rendant difficile l'interprétation des résultats. Ce choix a permis la construction de clusters plus homogènes en termes de comportement et donc de faciliter leurs descriptions.

Par ailleurs, par rapport à ma variable d'intérêt, le type d'établissement scolaire n'était pas associé de manière statistiquement significative avec la symptomatologie dépressive. Dans la littérature, le type d'établissement scolaire n'est pas considéré en soi comme un facteur de risque de la symptomatologie dépressive : par contre l'âge et la puberté sont des marqueurs connus dans la littérature de la symptomatologie dépressive (142,143).

Une des hypothèses pour expliquer l'importance majeure de cette variable pourrait être la suivante. Les différents établissements ne sont pas fréquentés par le même type d'adolescents (âge, genre, milieu socio culturel). Nous pouvons émettre l'hypothèse que le processus d'adolescence pourrait être différent. A la fois sur les moments (âge), et sur les mécanismes de construction de ce processus. Par exemple, l'orientation en lycée professionnel ou agricole en France, peut être vécue comme une forme d'exclusion par rapport au cycle « normal ». L'école est alors perçue comme celle qui exclue. Il y a donc une construction de l'expérience scolaire différente pour les élèves de ces différents établissements. Construction qui pourrait influer sur le processus d'adolescence.

Parmi les 39 variables différentes sélectionnées comme particulièrement importantes dans la structure des clusters, 27 étaient communes aux filles et aux garçons, quatre étaient spécifiques des garçons et huit spécifiques des filles. Les variables investiguant le redoublement, le sport, ou bien la participation à des jeux dangereux, montraient un poids particulièrement important dans les clusters

féminins et contribuaient plus faiblement à la structure des clusters masculins. Néanmoins sur ces variables, les différences de poids inter genre étaient relativement faibles : participation à des jeux dangereux ($poids_{filles}$=0,97, $poids_{garçons}$=0,89), faire du sport régulièrement ($poids_{filles}$=1, $poids_{garçons}$=0,89), redoublement ($poids_{filles}$=0,99, $poids_{garçons}$=0,87). Le poids de sélection de ces variables chez les garçons était largement supérieur à la moyenne de 0,74 (E-T=0,29). De plus, les différences de prévalence entre clusters « à haut risque » et clusters « à bas risque » corroboraient une certaine importance de ces variables dans les profils de risque des garçons (e.g le taux de participation à des jeux dangereux variait entre 2,5% et 11,9% dans les clusters à « bas risque » et entre 13,1% et 31,5% dans les clusters à « haut risque » chez les garçons).

Réciproquement, chez ces derniers, *« parler le plus facilement de l'actualité avec ses parents qu'avec d'autres »,* est une des variables particulièrement importantes dans la structure des clusters.

L'analyse par « régression sur profil », a permis d'analyser conjointement l'ensemble des variables explicatives et interpréter plus facilement les combinaisons (ou interactions). Dans ce chapitre de thèse, cette méthode a entre autres, permis de déterminer « avec finesse » des profils d'adolescents. Des différences de comportements apparaissaient au sein même de clusters à « haut risque ».

Par exemple, chez les filles, la pratique d'un sport était une variable majeure de la structuration des clusters. Au collège, quasiment toutes les filles du cluster à « haut risque » pratiquaient un sport. En revanche, au lycée, qu'il soit général et technologique, professionnel ou agricole, la relation entre la pratique d'un sport et le niveau de symptomatologie dépressive est plus complexe. En effet, parmi les deux clusters à « haut risque », des différences apparaissaient dans la prévalence en lycée général et technologique. Les filles du cluster le plus à risque (i.e cluster 18) étaient parmi les moins sportives (32%), tandis que celles du cluster 17 faisaient parties des plus sportives (98%). En lycée professionnel ou agricole, les filles du cluster le plus à risque pratiquaient peu de sport.

A titre d'exemple, chez les garçons, quasiment tous ceux des clusters à « haut risque » consommaient des substances psychoactives. Les prévalences de consommation (intensive ou expérimentation) étaient systématiquement les plus les élevées sauf pour un cluster. Ce cluster (cluster 14), était le cluster le plus à risque de tous et regroupait des garçons qui ne consommaient pas du tout

Cette méthode supervisée a pour avantage principal d'utiliser à la fois la variable d'intérêt clinique et la structure des données pour effectuer le partitionnement permettant ainsi de regrouper dans les clusters des sujets ayant des caractéristiques similaires en termes de facteurs de risque. La participation de la variable d'intérêt clinique à la constitution des clusters était un atout majeur en comparaison avec les méthodes de classification automatique non supervisées (e.g K-means, classification ascendante hiérarchique, classes latentes), où les clusters sont construits uniquement

sur les variables explicatives et l'association entre les clusters et la variable d'intérêt est analysée *a posteriori*. De plus, la « régression sur profil » a pour avantage de permettre d'analyser simultanément de nombreuses variables explicatives fortement corrélées.

Néanmoins, l'une des principales limites de cette méthode demeure dans le calcul du critère de sélection de variable. En effet, le calcul du poids de sélection, prend peu en compte la variable d'intérêt clinique ; il dépend principalement de la distribution de la variable explicative au sein des clusters. En d'autres termes, le poids de sélection latent, favorise les variables avec des différences inter clusters marquées. Les variables avec des différences inter clusters moins importantes, mêmes très associées à la symptomatologie dépressive sont ainsi défavorisées. Par exemple, le poids de sélection latent de l'expérimentation d'alcool avec du cannabis est de 1 chez les filles ; reflétant des prévalences très différentes selon les clusters variaient de 0% à 76,36%.

De plus, il n'existe pas de consensus sur le seuil de poids de sélection à choisir (135). Pour rappel, dans le cadre de ma thèse, une variable a été considérée comme majeure si son poids de sélection latent était supérieur ou égal à 0,95. Papathomas et al., ont utilisé dans leur étude un seuil à 0,75 (136). Le choix du seuil n'a pas de signification clinique ou statistique, il est effectué de façon pragmatique. En d'autres termes, certaines variables explicatives peuvent avoir contribuées de façon importante au partitionnement mais le seuil choisi les a exclus de l'analyse.

En conclusion, la « régression sur profil » s'est avérée intéressante sur le plan méthodologique en mettant en évidence des profils contrastés d'adolescents à risque de présenter une forte symptomatologie dépressive. La principale limite de cette approche est liée au fait que la modélisation conjointe de la variable d'intérêt clinique et des covariables peut être dominée par les termes spécifiques des covariables lorsqu'il y en a un grand nombre (144). Différentes techniques de sélection de variables ont été développées afin d'améliorer la sélection des variables majeures dans l'analyse par « régression sur profil » (145).

6. Discussion générale

La discussion de mon travail de thèse s'articule dans un premier temps autour des points forts et des limites de l'enquête « Processus d'adolescence ». Dans un second temps, elle reprend les principaux résultats des deux derniers objectifs de thèse (chapitres 4 et 5) et les discute. Enfin la discussion devient plus globale s'ouvrant sur les perspectives offertes par ce travail.

6.1. Points forts et limites de l'enquête « Processus d'adolescence »

L'utilisation dans ce travail de l'enquête « Processus d'adolescence » a de nombreux points forts mais aussi des limites, qui sont à la fois des avantages et des inconvénients. Cette enquête avait pour objectif d'étudier l'adolescent dans sa globalité et non de les partitionner selon leurs comportements. S'intéresser au processus dans sa globalité permet d'améliorer la connaissance sur cette génération mais aussi d'identifier de nouveaux indicateurs de difficultés dans le processus adolescent. Elle a fourni des données riches sur l'adolescent d'aujourd'hui. En premier lieu, elle a inclus 15235 adolescents, ce qui rend sa taille d'échantillon comparable aux grandes enquêtes en milieu scolaire (ESPAD, ESCAPAD). Deuxièmement, les données recueillies ont balayé des dimensions multiples, allant de la scolarité, à la consommation de substances psychoactives, l'alimentation, les loisirs, le sport, les jeux vidéo, en passant par la santé. Cependant, elle n'était pas spécifiquement prévue pour répondre à mon objectif de thèse ; qui était l'analyse des facteurs et marqueurs de risque de la symptomatologie dépressive. C'est pourquoi, il est important de noter que certaines variables explicatives connues dans la littérature n'ont pas été mesurées, par exemple la violence, le harcèlement scolaire ou les antécédents psychiatriques de la famille. Ici, l'enquête « Processus d'adolescence » apporte de nombreuses données mesurées concomitamment dans des champs différents alors que les autres études publiées sur les adolescents ont majoritairement axé leurs recherches sur des champs spécifiques. Ces champs ont été investigués plus fréquemment par de nombreuses questions/items.

Dans cette étude, la seule échelle de mesure d'intensité des symptômes dépressifs validée en français chez l'adolescent (ADRS) a été inclue. Elle a été créée dans les années 2000 devant l'absence d'outils de mesure de la dépression spécifique à l'adolescence (18). La création de l'ADRS a suivi les étapes référentes de validation des outils de mesure. En effet, une première phase de construction s'est appuyée sur des entretiens de recherche en méthode qualitative avec des adolescents et des psychiatres d'adolescents. L'échelle a été construite à partir du vocabulaire que les adolescents avaient utilisé pour décrire leur expérience dépressive. Cette expérience dépressive a pu être organisée en trois facettes, secondairement validées par les psychiatres d'adolescents.

Les seuils validés de cette échelle permettent de distinguer les adolescents à faible ou forte symptomatologie dépressive. Ces critères sont suffisamment proches pour qu'une forte symptomatologie dépressive telle que définie par l'ADRS puisse être considérée comme un diagnostic d'épisode dépressif caractérisé selon le DSM-V (10). Toutefois, il ne s'agit pas d'une échelle diagnostic, elle ne peut donc pas remplacer un entretien clinique structuré avec un clinicien. Cet auto questionnaire fréquemment utilisé dans les enquêtes en milieu scolaire se présente comme l'outil de référence pour mesurer la symptomatologie dépressive spécifique à l'adolescence.

En ce qui concerne les autres dimensions de l'enquête, il existe de nombreuses échelles utilisées en population adolescente qui évaluent les autres champs de l'enquête « Processus d'adolescence » et qui auraient pu être utilisées. Par exemple, l'échelle auto évaluative CAST (Cannabis Abuse Screening Test) qui permet la détection des usages problématiques de cannabis et validée en population adolescente en France, à travers le dispositif d'enquêtes nationales ESPAD (146). De même, l'échelle épidémiologique de mesure des usages problématiques de jeux vidéo chez les adolescents : La Game Addiction Scale (GAS), dite échelle de Lemmens est une des seules échelles de mesure de l'addiction aux jeux vidéo, qui ait été validée pour les adolescents (147). Les promoteurs de l'enquête ont fait le choix de ne pas inclure ces échelles dans le questionnaire, au profit de questions sur la consommation au cours de la vie, dans l'année et dans le mois de substances psychoactives. Ces questions étaient tirées du questionnaire de l'enquête ESPAD et ESCAPAD.

L'enquête « Processus d'adolescence » a interrogé des adolescents scolarisés « sur des territoires géographiques contrastés, se situant dans des contextes sociodémographiques très différents ». Les territoires concernés étaient : le département du Val de Marne (tirage au sort stratifié de collège, lycée général et technologique et lycée professionnel), la région Hautes-Alpes (ensemble des collèges et lycées généraux et technologique et lycées professionnels) et la région Poitou-Charentes (ensemble des collèges et lycées agricoles). Ils regroupent des réalités de distribution de richesse et d'urbanisme différentes (urbains, montagnards et ruraux). Cette enquête a inclus des établissements agricoles. A ma connaissance, les adolescents scolarisés dans ce type d'établissement, sont très rarement approchés par les autres enquêtes nationales en milieu scolaire. Néanmoins, ils sont inclus dans des enquêtes régionales très ciblées sur un comportement en particulier. La présence de cette population spécifique dans mon échantillon est pourvoyeuse d'hétérogénéité. Compte tenu des objectifs de l'enquête « Processus d'adolescence », il me semble que l'inclusion de ces adolescents était un atout majeur. L'objectif de cette inclusion était donc d'obtenir des données sur leurs comportements globaux, mais également de leur donner la parole. Les résultats obtenus ne sont également pas extrapolables à l'ensemble des adolescents, notamment aux adolescents non scolarisés

L'enquête « Processus d'adolescence » a peu d'équivalent parmi l'ensemble des enquêtes adolescentes pré existantes en milieu scolaire (ESPAD, ESCAPAD, HBSC). Bien que datant de 2013, il s'agit d'une des enquêtes françaises les plus récentes s'intéressant à l'adolescent dans sa globalité. La précédente date de 1993, elle avait été dirigée par Choquet et al., « Enquête Nationale » (1). La plupart des enquêtes en milieu scolaire sur cette population ne s'intéresse qu'à un certain type de comportement et recueillent des données ciblées : HBSC s'intéresse principalement au vécu scolaire, à la santé et à l'expérimentation et consommation de produits psychoactifs ; ou bien les enquêtes ESPAD et ESCAPAD qui s'intéressent essentiellement aux consommations de substances psychoactives. Malgré des similitudes sur le mode de recueil des données (reposant sur une méthodologie standardisée fréquemment utilisée dans les enquêtes en milieu scolaire) et un ensemble d'items identiques recueillis, la comparaison avec les résultats obtenus par ces enquêtes reste difficile, compte tenu des différences de types de populations inclus.

Enfin, cette enquête est transversale. L'ensemble des résultats obtenus sont donc à interpréter avec précautions. En effet, la nature de l'enquête ne permet pas d'appréhender la temporalité de l'association entre les variables explicatives et la symptomatologie dépressive. Seule une étude longitudinale aurait permis d'étudier le lien de causalité entre les facteurs et la variable d'intérêt (symptomatologie dépressive).

6.2. Discussion des principaux résultats

L'utilisation du LASSO, des méthodes d'agrégation d'arbres et de la « régression sur profil » ont mis en exergue des variables importantes dans la caractérisation d'une forte symptomatologie dépressive. Dans la suite de ce chapitre, j'ai choisi de ne discuter que les variables qui à mon sens sont importantes dans l'interprétation clinique.

- La participation à des « jeux dangereux »

Dans l'analyse des variables explicatives de la symptomatologie dépressive, il est apparu dans les modèles d'agrégation d'arbres (SGD et RF) et l'analyse par « régression sur profil », que la participation à des « jeux dangereux » était associée à une forte symptomatologie dépressive en particulier chez les filles.

Dans l'enquête « Processus d'adolescence », 9,26% des adolescents ont déclaré avoir déjà participé à des jeux dangereux (5,2% des filles et 13,7% des garçons). Ces prévalences sont cohérentes avec celles de l'enquête enquête TNS-SOFRES menée en 2007, qui a évalué les taux de participation à des « jeux dangereux » avec cette appellation chez les enfants et les adolescents. Dans cette enquête, 12% des 489 jeunes interrogés (âgés de 7 à 17 ans), déclaraient avoir déjà participé à un jeu dangereux (148).

Dans une étude transversale en Gironde sur 832 collégiens de la sixième à la troisième interrogeant les trois types de jeux, 10% des collégiens ont déclaré avoir déjà joué à un jeu d'asphyxie, 26,9% à un jeu d'agression, 26,3% à un jeu de défi. D'une manière générale, les données épidémiologiques sont rares et concernent essentiellement les jeux de non oxygénation.

Très peu d'études ont évalué le lien entre la participation à des « jeux dangereux » et les symptômes dépressifs. Ces études se focalisaient sur les pratiques d'auto ou hétéro-asphyxie. Deux études françaises transversales réalisées en 2009 et 2013 sur un total de 1771 collégiens de la sixième à la troisième ont mis en évidence une association entre, la participation à un jeu d'étouffement et la présence de symptômes dépressifs (33). D'autres études ont montré une association significative entre la participation à des jeux de non-oxygénation et des facteurs de risque, tels que la consommation de substances psychoactives, des comportements sexuels à risque, des conduites auto mutilatoires. Ces études s'intéressaient également au lien entre la pratique d'un sport à risque, les rapports sexuels forcés, ou les comportements suicidaires (e.g, idées suicidaires, et avoir déjà fait une tentative de suicide) et la participation à des jeux dangereux (149–152).

Mon travail de thèse est la seule étude à ma connaissance ayant évalué les différences inter-genre quant à la force de l'association entre la participation à des « jeux dangereux » et la symptomatologie dépressive. La participation à des « jeux dangereux » apparait comme un indicateur important dans la caractérisation des adolescents à risque de présenter une forte symptomatologie dépressive, en particulier chez les filles.

Différentes hypothèses pourraient expliquer l'association entre la participation à des « jeux dangereux » et les symptômes dépressifs. Chez les adolescents présentant des troubles dépressifs, les jeux dangereux seraient utilisés comme une stratégie d'automédication pour contrebalancer des émotions négatives découlant d'affects dépressifs douloureux (34). Selon Bernadet et al., le choix du type de jeu dépendrait d'un profil psychologique spécifique à chaque jeu. Les adolescents impulsifs et dépressifs se dirigeraient plutôt vers des jeux de non-oxygénation et des jeux d'agression. Ceux présentant une forte recherche de nouveauté se dirigeraient plutôt vers des jeux de défis et une diversification des types de jeux dangereux (153). Réciproquement, les jeux dangereux auraient des répercussions durables sur le développement psycho-affectif de l'enfant ou de l'adolescent, avec notamment des manifestations de dépréciation de soi, des troubles du sommeil, des phobies scolaires, et des troubles anxieux et dépressifs (37).

Les résultats obtenus sont à interpréter avec précautions. La question évaluant la participation à des « jeux dangereux » dans l'enquête « Processus d'adolescence » est une question généraliste pouvant ne pas recouvrir les pratiques de jeux dangereux habituellement évaluées dans les études (paragraphe 1.1.3 page 18).

- Consommation de substances psychoactives

Les variables explicatives de consommation de substances psychoactives licites ou illicites (expérimentation, consommation intensive au cours des 30 derniers jours) et spécifiquement pour l'alcool, l'ivresse et le contexte de consommation d'alcool, apparaissaient comme significatives en analyse bivariée chez les filles comme chez les garçons.

Ces variables ont été identifiées comme majeures dans les modèles d'agrégation d'arbres (dix premières variables des modèles SGD et RF) et l'analyse par « régression sur profil ». Dans l'analyse des modèles d'agrégation d'arbres, les variables de consommation identifiées étaient uniquement des variables de consommation d'alcool : *« Quand vous buvez c'est plutôt seul »* chez les filles et chez les garçons ; de consommation intensive d'alcool au cours des 30 derniers jours, *« Quand vous buvez c'est plutôt en soirées ou avec des copains »*, d'intensité de l'ivresse, uniquement chez les garçons. Aucune de ces variables n'étaient identifiées parmi les dix premières les plus importantes chez les filles (uniquement sélectionnées dans le modèle LASSO Tableau 10). Chez les garçons, elles représentent 18,75% (3/16) des variables les plus importantes dans au moins un des deux modèles d'agrégation d'arbres (16 variables différentes parmi les 10 variables les plus importantes des deux modèles Tableau 11). Dans l'analyse par « régression sur profil », les variables de consommation intensive et d'expérimentation de substances psychoactives, représentaient quant à elles, 45% (16/35) des variables majeures chez les filles et 54% (16/31) chez les garçons (Annexe 4, page 171).

Mes résultats sont en adéquation avec les études qui se sont intéressées au lien entre dépression et la consommation de substances psychoactives. En effet, la revue de la littérature effectuée par Cairns et al. 2014, a mis en évidence que la consommation d'alcool (fréquence et quantité), de tabac, de cannabis, d'autres drogues illicites mais aussi la poly consommation étaient associées à un niveau de dépression élevé.

Cependant, dans mon travail de thèse comme dans la revue de la littérature effectuée par Cairns et al., la direction de l'association ne peut être établie (30). Il est plausible que la relation soit bidirectionnelle : c'est-à-dire que la dépression peut conduire à la consommation de substances ou que la consommation de substances peut être considérée comme une forme d'automédication de la dépression. Concernant la poly consommation, la revue de Cairns et al., et les résultats obtenus dans

ma thèse, ne permettent pas de déterminer quel type de poly consommation avait un impact sur la dépression à l'adolescence.

- Temps de jeux vidéo

Dans l'enquête « Processus d'adolescence », le temps passé à jouer aux jeux vidéo (en semaine, week-end et vacances) par jour était associé à une forte symptomatologie dépressive et a été identifié parmi les variables majeures dans les modèles d'agrégation d'arbres (2/14 chez les filles et 3/16 chez les garçons), ainsi que dans l'analyse par « régression sur profil » (3/35 chez les filles et 3/31 chez les garçons). Les adolescents des clusters à « haut risque », chez les filles comme chez les garçons, déclaraient les prévalences de temps passés à jouer plus de 3 heures à des jeux vidéo étaient parmi les plus élevées de toutes les prévalences à établissement scolaire équivalent.

Les conclusions des études examinant l'association potentielle entre le temps passé à jouer aux jeux vidéo et le risque de dépression chez les adolescents sont équivoques. La plupart des études ont utilisé des modèles de régression pour tester une association entre le temps moyen par jour et la dépression ; certaines ont fait état d'une association positive significative (54), d'autres ont suggéré des associations négatives (55) ou non significatives (154–156).

Mes résultats sont en adéquation avec l'étude de Maras et al., montrant une association positive significative entre le temps passé à jouer aux jeux vidéo et la dépression à l'adolescence (54).

Cependant, les prévalences des clusters à « haut risque » n'étaient pas plus hautes que les autres clusters. Comment expliquer ce résultat ? Une première hypothèse serait que les filles jouant à des jeux vidéo ont des profils très différents de celles ne jouant pas aux jeux vidéo. Une deuxième hypothèse serait que des facteurs de confusion interviennent dans la relation entre le temps de jeux vidéo et la dépression. L'étude de Maras et al., a mis en évidence que l'IMC ou encore l'activité physique interviendrait dans cette relation (54).

- Sport

Dans l'enquête « Processus d'adolescence », la pratique d'un sport était liée à la symptomatologie dépressive quel que soit le genre : les filles pratiquaient du sport de loisir tandis que les garçons principalement du sport de compétition.

Les méthodes d'agrégation d'arbres ont mis en évidence que la pratique régulière d'un sport était associée à une forte symptomatologie dépressive chez les filles comme chez les garçons (variable majeure). Cette variable explicative a également été majeure pour la structuration des clusters dans l'analyse par « régression sur profil » mais uniquement chez les filles.

Plusieurs études se sont intéressées au lien entre la pratique d'une activité physique et la santé mentale (plus particulièrement anxiété et dépression). Une comparaison avec ces études est intéressante étant donné que le sport fait partie de l'activité physique par définition. Les études de Eime et al., et Johnson et al., (48,157), ont mis en évidence que la pratique d'une activité physique était associée à un bas niveau de dépression. Malgré la différence de notion évaluée (pratique d'une activité physique, le plus souvent évaluée *via* une échelle), mes résultats sont en contradiction avec la littérature, plus particulièrement au collège. Les filles des clusters à « haut risque » pratiquaient beaucoup de sport.

Il existe également des études qui se sont intéressées tout comme moi à la pratique d'un sport *via* un item (par exemple « au cours de la dernière semaine, combien de fois avez-vous participé à un sport de type baseball etc… ? »). Ces études ont mis en évidence que la pratique d'un sport était associée à un bas niveau de dépression (51,53). Mes résultats au collège et dans un cluster de lycée général et technologique chez les filles sont encore une fois en contradiction avec la littérature. Au collège, la prévalence de pratique régulière d'un sport était la plus élevée dans le cluster à « haut risque » (92%). Au lycée professionnel ou agricole, les filles pratiquaient peu de sport dans le cluster à « haut risque » (51%). Au lycée général et technologique, les résultats sont plus complexes à interpréter : parmi les deux clusters à « haut risque », les filles d'un cluster à « haut risque » pratiquaient peu de sport (cluster le plus à risque, 32%) tandis que les filles de l'autre cluster à « haut risque » pratiquaient beaucoup de sport (97%).

De même que pour le temps passé à jouer aux jeux vidéo, ces contradictions pourraient être expliquées par l'intervention de facteurs confondants dans l'association étudiée.

- Influence de la pratique d'un sport et du temps passé à jouer à des jeux vidéo sur la symptomatologie dépressive

Une étude transversale canadienne s'est intéressée à l'interaction entre le temps passé à jouer à des jeux vidéo, la pratique d'une activité physique et la dépression à l'adolescence (54). Les modèles de régression multiples de Maras et al., ont montré que le temps passé à jouer à des jeux vidéo était associé à la sévérité de la symptomatologie dépressive en ayant contrôlé entre autres sur l'âge, le sexe, l'activité physique, l'IMC, et la durée du temps passé à regarder la télévision ou à utiliser un ordinateur.

La relation entre le temps passé à jouer à des jeux vidéo, la symptomatologie dépressive et la pratique d'un sport a pu être étudiée à partir des résultats de l'analyse par « régression sur profil » chez les filles. Il est apparu dans mes résultats qu'au collège, les adolescentes des clusters à « haut risque» étaient parmi les plus sportives et joueuses de jeux vidéo. Au lycée général et technologique, certaines filles des clusters à « haut risque », pratiquaient beaucoup de sport et jouaient modérément aux jeux

vidéo ; d'autres pratiquaient peu de sport et jouaient peu aux jeux vidéo. Au lycée professionnel ou agricole, les filles des clusters à « haut risque », pratiquaient peu de sport et jouaient modérément aux jeux vidéo.

Boers et al., ont émis une hypothèse sur le lien entre le temps passé devant un écran, la pratique d'une activité physique et la santé mentale. Il s'agit de l'hypothèse du « déplacement ». Cette hypothèse suggère que le temps consacré aux activités sur écran peut remplacer le temps consacré à des activités plus productives et/ou actives, notamment l'activité physique ou les communications interpersonnelles, et peut ainsi avoir un impact sur la santé mentale des jeunes (155).

De nombreuses hypothèses ont été avancés pour cette expliquer cette relation. Certains jeunes qui consacrent beaucoup de temps aux jeux vidéo s'isolent socialement. Ils ont plus de problèmes de sommeil, ce qui pourrait compromettre leur capacité à faire face au stress, entraînant un sentiment accru de dépression ou d'anxiété. Enfin, le temps passé devant un écran peut également remplacer le temps passé à faire une activité physique, ce qui est préoccupant compte tenu des conclusions précédentes de la littérature selon lesquelles l'activité physique est associée à une diminution des symptômes d'anxiété et de dépression. Ce qui est contraire aux résultats trouvés dans mon analyse où la pratique d'un sport était associée à une forte symptomatologie dépressive au collège et au lycée générale et technologique.

- Influence de la pratique régulière de sport et de la consommation de substances psychoactives sur la symptomatologie dépressive.

Pour rappel, il est apparu que les collégiennes du cluster à « haut risque » étaient parmi les plus consommatrices de substances psychoactives et les plus sportives. Elles déclaraient également pratiquer un sport à risque. Au lycée général et technologique, les deux clusters à « haut risque », présentaient des caractéristiques différentes quant à la pratique d'un sport. En effet, un cluster regroupait des filles qui pratiquaient beaucoup de sport et consommaient beaucoup de substances psychoactives. L'autre cluster à « haut risque » regroupait quant à lui, des filles qui pratiquaient peu de sport et consommaient beaucoup de substances psychoactives. Ce profil était identique dans le cluster à « haut risque » de lycée professionnel ou agricole. À la vue de ces résultats, on peut supposer que la pratique d'un sport et la consommation de substances psychoactives jouent un rôle conjoint sur la forte symptomatologie dépressive. On remarque également que ces filles déclaraient pratiquer un sport à risque.

Dans la littérature, la pratique d'un sport est le plus souvent associée à un faible risque de présenter une dépression, tandis que la consommation de substances est associée à risque élevé de présenter une dépression à l'adolescence.

L'étude de Choquet et al., sur l'activité sportive à l'adolescence et les troubles des conduites associées, a mis en évidence que les filles qui pratiquaient beaucoup de sport (plus de 8 heures), adoptaient plus de comportements à risque (consommation de substances, troubles alimentaires…) (1). L'hypothèse sous-jacente était qu'à travers le sport et la consommation de substances, elles mettaient leur corps à l'épreuve. Dans mon travail, pour les clusters à « haut risque », on retrouve ces résultats, chez les collégiennes, les filles en lycée professionnel ou agricole et chez une partie des filles en lycée général et technologique. Il serait intéressant d'étudier le type de pratique et les conditions de cette pratique. Existe-t-il un mécanisme de la « modération » tant pour la pratique sportive que pour les troubles et comportements ?

- Différences filles/ garçons

La « régression sur profil » a permis de mettre en évidence des combinaisons de variables différentes afin de partitionner les adolescents dans des sous-groupes selon le risque de présenter une forte symptomatologie dépressive.

Vingt-sept variables étaient communes aux deux genres et des similarités de comportements ont été observées entre les deux genres. Les filles et les garçons des clusters à « haut risque », avaient plus tendance à avoir déjà eu des rapports sexuels, avoir eu un premier rapport sexuel à risque et à consommer des substances psychoactives, sauf pour le cluster le plus à risque chez les garçons.

En effet, les garçons du cluster 14, cluster le plus à risque de présenter une forte symptomatologie dépressive, ne consommaient pas de substances psychoactives. Il s'agissait d'un cluster mixte, avec à la fois des collégiens et des lycéens (52,46% de collégiens), scolarisés majoritairement dans le Val de Marne. Avec les données disponibles dans l'enquête « Processus d'adolescence » et les variables majeures extraites de mon analyse par « régression sur profil », il n'était pas possible de distinguer ce cluster à « haut risque » des autres clusters masculins.

- Ressenti vis-à-vis de l'école

Le ressenti vis-à-vis de l'école était évalué dans l'enquête « Processus d'Adolescence » par une question simple « Actuellement que pensez-vous de l'école ? ». En analyse bivariée, cette question était associée significativement avec la symptomatologie dépressive chez les filles comme chez les garçons. Elle a également été inclue dans l'analyse des méthodes d'agrégation d'arbres et la « régression sur profil ». Les poids de sélection latent de l'analyse par « régression sur profil » sur cette variable était inférieur à 0,70 dans les deux genres ($poids_{filles}$=0,68, $poids_{garcons}$=0,61). Pour les filles déclarant aimer l'école dans l'ensemble de l'échantillon, les fréquences variaient de 46,57% à 89,42% (avec une fréquence variant de 46,57% à 58,18% dans les clusters à « haut risque », parmi les plus

faibles fréquences). Chez les garçons cette fréquence variait de 35,39% à 75,73% (avec une fréquence variant de 35,39% et 53,48% également parmi les plus faibles fréquences).

A ma connaissance en France, seule l'enquête HBSC s'est intéressée à cette question du ressenti vis-à-vis de l'école comme facteur de risque, en étudiant son lien avec les symptômes dépressifs. Toutefois dans cette enquête la mesure ne se fait pas avec un seul item mais avec une échelle validée. Dans l'enquête « Processus d'adolescence », il n'y avait qu'un seul des items de l'échelle, item utilisé comme indicateur de bien être à l'école mais devant être plutôt considéré comme une expression symptomatique plutôt qu'un facteur de risque.

7. Conclusions et perspectives

Cette thèse a abordé la question de l'apport des méthodes de « fouille de données » à l'épidémiologie psychiatrique et plus particulièrement à la dépression de l'adolescent. J'ai utilisé des approches

variées, pour étudier l'association entre la symptomatologie dépressive à l'adolescence et ses variables explicatives, afin de déterminer des profils d'adolescents à risque de présenter une forte symptomatologie dépressive.

Dans ce travail de thèse, la cartographie de l'application des méthodes de DMML en santé publique et épidémiologie, a permis de montrer que leur utilisation était relativement récente, en particulier en épidémiologie psychiatrique. Au début du travail sur mon premier objectif de thèse (2016), très peu de recherche en épidémiologie psychiatrique utilisaient du DMML. Cette tendance a évolué, et on assiste maintenant à une utilisation plus importante de ces méthodes.

Durant toute ma thèse (de janvier 2016 à décembre 2020), j'ai suivi l'évolution des méthodes de DMML et des programmes permettant de les utiliser. Par exemple, au début de ma thèse, il existait de nombreuses librairies différentes, permettant, d'effectuer du DMML (i.e librairie « glmnet » utilisé par exemple pour effectuer des régressions LASSO, librairie « RandomForest », utilisé pour effectuer des forêts aléatoires). En 2018, le package « Caret » a été créé, il regroupe plusieurs méthodes de DMML couramment utilisées je l'ai donc utilisé dans ma thèse. De par leur caractère innovant, les méthodes et les algorithmes de DMML ne cessent d'évoluer. Ce perpétuel remaniement a engendré de nombreuses discussions avec mes directeurs de thèse afin d'optimiser les différents paramètres nécessaires à la réalisation de mes travaux. L'un de mes objectifs de thèse consistait en l'application des méthodes de DMML sur l'enquête « Processus d'Adolescence ». A ce moment-là de ma thèse, très peu d'études utilisant des données similaires à celles utilisées dans ma thèse, ont fait état de l'utilisation des méthodes de DMML pour répondre à leur objectif. Sur la base de la littérature existante, des choix méthodologiques ont dû être réalisés, par exemple sur le nombre d'itération, le choix des hyperparamètres à tester etc…

L'hypothèse de départ de cette thèse était que l'utilisation de techniques issues du DMML pour prédire la dépression à l'adolescence serait meilleure que les modèles de régression (ici régression pénalisée LASSO). Cette hypothèse est née du constat que de par leur caractère innovant, les méthodes de DMML prédirait mieux et aiderait à l'interprétation des associations complexes entre la symptomatologie dépressive et les variables explicatives. Comme nous l'avons vu dans la description des résultats, ces derniers ne permettent pas de l'affirmer.

Malgré de faibles performances prédictives, les méthodes d'agrégation d'arbres (RF, SGD) ont montré de bonnes performances dans la sélection de variables. Elles ont permis de sélectionner les variables les plus importantes dans la prédiction de la symptomatologie dépressive. Toutefois, ces méthodes sont par définition difficilement interprétables. C'est pourquoi, l'utilisation d'une méthode de partitionnement (la « régression sur profil ») a été un atout dans cette thèse. En effet, elle permet de

trouver et d'explorer des combinaisons de variables afin d'en extraire des connaissances pertinentes sur les sous-groupes. Toutefois, certains profils restent difficilement interprétables (e.g cluster 14 chez les garçons, le cluster le plus à risque ; à partir des données disponibles dans l'enquête « Processus d'Adolescence », il n'était pas possible de le distinguer des autres profils à haut risque en termes de comportements).

Les deux méthodes (classification et partitionnement) ont mis en évidence les mêmes familles de variables. Au vu de ces résultats, les méthodes de DMML devraient être utilisées de manière complémentaire à la régression logistique. Pour une problématique similaire, les méthodes de DMML pourraient par exemple, être dans un premier temps utilisées pour identifier des variables majeures dans la prédiction d'une variable d'intérêt clinique pour ensuite utiliser ces variables dans un modèle de régression.

D'un point de vue méthodologique, l'application des méthodes de DMML a permis d'identifier des variables majeures qui se retrouvent être celles qui sont fréquemment retrouvées par d'autres techniques et étudiées dans la littérature. Il est donc intéressant de se demander comment tenir compte des résultats obtenus *via* ces différentes méthodes, dans la pratique clinique et ainsi faire le lien entre les deux valences de ma thèse (méthodologique et clinique).

Dans cette réflexion, on peut s'appuyer sur le guide, « HEADSS » proposé en 1988, par Goldenring aux cliniciens. Ce guide d'entretien clinique est destiné à orienter les cliniciens dans l'évaluation de la santé psychosociale des adolescents. Il a proposé aux cliniciens d'interroger les adolescents sur sept sphères : i) l'environnement familial, ii) l'école, iii) les activités, iv) la consommation de substances psychoactives, v) la sexualité, vi) le suicide, vii) la sécurité. Suite à des discussions avec plusieurs pédopsychiatres, cet outil semble peu utilisé en pratique.

Toujours dans l'optique du lien entre la valence méthodologique et clinique, l'objectif secondaire de mon travail de thèse était de pouvoir proposer aux cliniciens des indicateurs simples pouvant induire des pistes de repérage d'adolescents à risque de présenter une forte symptomatologie dépressive. Force est de constater que les variables identifiées comme majeures dans mon analyse et ce, grâce à l'application des différentes méthodes font parties des sphères identifiées dans ce guide.

En effet, on retrouve des variables s'intéressant à l'école, la consommation de substances psychoactives, la pratique d'un sport, le temps passé à jouer à des jeux vidéo, la participation à des jeux dangereux, les troubles du rythme du sommeil. Plusieurs sphères de ce guide sont d'ailleurs bien documentées dans la littérature scientifique (consommation de substances psychoactives, sexualité, suicide, activité), d'autres le sont moins (école et sécurité). Même si ces variables sont déjà ancrées dans une certaine pratique clinique, il semble pertinent de porter une attention particulière sur trois

indicateurs en particulier: la participation à des jeux dangereux, le ressenti vis-à-vis de l'école et la pratique d'un sport. Dans l'enquête « Processus d'adolescence », sur l'ensemble de la population, 12,1% des adolescents présentaient une forte symptomatologie dépressive. Parmi tous les adolescents, 23,1% de ceux participant à des jeux dangereux présentaient une forte symptomatologie dépressive, 19,4% déclarant ne pas aimer l'école présentaient une forte symptomatologie dépressive et 22,6% qui pratiquaient régulièrement un sport, présentaient une forte symptomatologie dépressive. Ces prévalences non négligeables montrent que la prise en compte de ces indicateurs dans le dépistage des adolescents à forte symptomatologie dépressive est primordiale et peut conduire à une prise en charge spécifique.

Intéressons-nous tout d'abord à la sphère « Ecole ». En recherche comme en clinique, la sphère de l'école est souvent explorée par le versant phobie scolaire. La phobie scolaire, est un phénomène complexe évaluant des dimensions particulières, explorées sous différents angles, pas toujours simple à évaluer en entretien clinique. Le ressenti vis-à-vis de l'école, avec une question simple comme « aimez-vous l'école ? » serait un indicateur simple à explorer et pourrait aider le clinicien à repérer les adolescents potentiellement à risque de présenter une forte symptomatologie dépressive.

De même, la participation à des « jeux dangereux » pourrait davantage être explorée durant un entretien clinique (sphère sécurité). En recherche, la participation à des jeux dangereux est très peu étudiée. En clinique, interroger les adolescents sur leur participation à des jeux dangereux n'est pas toujours réalisé. Leur poser cette question simple de type « avez-vous déjà participé à des jeux dangereux ? » pourrait permettre aux cliniciens de repérer le profil d'un adolescent à risque de présenter une forte symptomatologie dépressive étant donné la prévalence non négligeable dans l'enquête (notamment chez les filles où cette variable est majeure).

En ce qui concerne la pratique d'un sport, et particulièrement la pratique d'un sport à risque, du côté de la recherche, il a été mis en évidence que la pratique d'un sport était un facteur de résilience de la dépression à l'adolescence et était un indicateur de bonne santé. Cependant, dans mon travail de thèse, la pratique d'un sport était associée à une augmentation du risque de présenter une forte symptomatologie dépressive chez une partie des filles. Lors d'un entretien clinique, s'intéresser à la pratique d'un sport voire d'un sport à risque et ce, notamment chez les filles, pourrait aider à mieux repérer celles, pouvant être à risque de présenter une dépression lorsque par exemple, elles consomment également des substances psychoactives.

Pour conclure, mon travail de thèse, a permis de montrer que l'application des méthodes de DMML est en augmentation durant la dernière décennie notamment en santé mentale. Ces méthodes en perpétuel remaniement sont à adapter à la problématique que l'on souhaite étudier. Malgré cette

augmentation, leur utilisation reste discutée. En effet, de plus en plus de débats émergent quant à l'intérêt de ces méthodes par rapport aux modèles de régression. Certains tendent à accorder la « toute puissance » à ces méthodes. D'autres au contraire, ont tendance à nier leur intérêt et leur efficacité notamment à cause de l'absence d'hypothèse *a priori*. Dans ce travail, malgré les faibles performances prédictives, les méthodes de DMML sont apparues comme de bonnes méthodes dans la sélection de variables majeures dans la prédiction du risque de présenter une forte symptomatologie dépressive. Elles ont mis en évidence des phénomènes complexes entre les variables explicatives (i.e la combinaison de la consommation de substances psychoactives et de la pratique d'un sport chez les filles à risque de présenter une forte symptomatologie dépressive). En d'autres termes, ce travail a permis de relativiser la pensée de certains sur la toute-puissance attribuée aux méthodes de DMML. Mais aussi de mettre en évidence que l'opposition entre ces deux philosophies n'est pas nécessaire. L'utilisation complémentaire des méthodes de DMML et des modèles de régression pourrait être une issue possible afin de mieux comprendre les associations entre des facteurs/ marqueurs dans le but de prédire une variable d'intérêt clinique. Les méthodes de DMML pourraient par exemple être utilisées dans un premier temps pour sélectionner des variables majeures parmi un grand ensemble de variables explicatives.

Annexe 1: Questionnaire de l'enquête "Processus d'Adolescence"

Inserm

Institut national
de la santé et de la recherche médicale

CENTRE HOSPITALIER
Fondation Vallée
Psychiatrie de l'Enfant et de l'Adolescent - Gentilly

Enquête Processus d'Adolescence

Cette enquête multicentrique est proposée aux adolescents entre 13 et 18 ans dans les Hautes-Alpes, en Poitou-Charentes et dans le Val-de-Marne. Sa réussite dépend de vous. Le questionnaire que vous avez accepté de remplir est **confidentiel** et **anonyme**. Vous pouvez donc y répondre en toute confiance. Par contre, il est important que vous le fassiez de manière sincère pour ne pas fausser les résultats.

Ce questionnaire n'est pas fait pour contrôler vos connaissances ou vous juger : il va servir à mieux comprendre les adolescents en général en recueillant des données nouvelles pour améliorer l'aide à leur apporter en cas de difficulté et leur proposer une prévention mieux adaptée.

Pour Répondre :
Pour chaque question, mettez une croix dans la case que vous choisissez : ☒ ☐
En cas d'erreur, noircissez complètement la case et inscrivez une croix dans la bonne case. Exemple: ■ ☒

Attention, pour certaines questions, une seule réponse est possible, pour d'autres *(texte en italique)* vous pouvez indiquer plusieurs réponses. Exemple :

4. Avez-vous un autre adulte proche de vous à qui parler de votre santé ?
(Plusieurs réponses possibles)

Votre père	☒ 1	*Un ami de votre famille*	☒ 8
Votre mère	☐ 2	*Votre CPE*	☐ 9

Quand vous voyez une ligne telle que celle-ci _____, vous devez écrire votre réponse en toutes lettres.

Il n'y a pas de bonnes ou de mauvaises réponses, choisissez uniquement la réponse qui vous ressemble le plus. Lisez attentivement chaque affirmation, mais ne passez pas trop de temps pour décider de la réponse. Répondez à toutes les questions, même si vous n'êtes pas très sûr(e) de la réponse.

Écrivez au **stylo noir uniquement et n'utilisez jamais de correcteur**. Les chiffres à côté des cases nous servent juste à coder la question, ils sont sans importance pour vous.

Si vous avez des difficultés pour répondre aux questions, vous pouvez solliciter l'adulte qui vous encadre et aussi l'inscrire à la fin du questionnaire sur la fiche « commentaires ».

Merci de votre participation

1 ▶▶▶

▶ QUELQUES RENSEIGNEMENTS SUR VOUS

1.1. **Quelle est votre année de naissance ?** ☐☐☐☐

1.2. **Êtes-vous :** Une fille ☐2 Un garçon ☐1

1.3. **En quelle classe êtes-vous ?**
(Une seule réponse possible)

4eme	☐1	Deuxième cycle agricole, 1re	☐11
3eme	☐2	Deuxième cycle agricole, Terminale	☐12
Seconde (générale ou technologique)	☐3	CAP, CAPA 1re année	☐13
Première (générale ou technologique)	☐4	CAP, CAPA, 2eme année	☐14
Terminale (générale ou technologique)	☐5	CAP, CAPA, 3eme année	☐15
Seconde professionnelle	☐6	BEP, BEPA, 1re année	☐16
Première professionnelle	☐7	BEP, BEPA, 2eme année	☐17
Terminale professionnelle	☐8	SEGPA (6eme, 5eme, 4eme, 3eme)	☐18
Premier cycle agricole (4eme, 3eme)	☐9	SEGPA (CAP)	☐19
Deuxième cycle agricole, 2nde	☐10	Autre	☐20

1.4. **Avez-vous déjà redoublé ?**
(Une seule réponse possible)

Jamais ☐1 **Une seule fois** ☐2 **2 fois ou plus** ☐3

1.5. **Actuellement que pensez-vous de l'école ?**
(Une seule réponse possible)

Je l'aime beaucoup	Je l'aime un peu	Je ne l'aime pas beaucoup	Je ne l'aime pas du tout
☐1	☐2	☐3	☐4

▶ VOTRE SANTÉ ET VOUS

2. Par rapport aux personnes de votre âge, diriez-vous que votre état de santé est :
(Une seule réponse possible)

Pas du tout satisfaisant	Peu satisfaisant	Plutôt satisfaisant	Très satisfaisant
☐1	☐2	☐3	☐4

3. La dernière fois que vous avez vu un médecin c'était
(Une seule réponse possible)

Au cours de la semaine dernière	Au cours du mois dernier	Au cours de l'année dernière	Avant l'année dernière
☐1	☐2	☐3	☐4

3.1. **Lorsque vous voyez un médecin, arrivez-vous facilement à lui parler de vos problèmes ?**

Oui ☐2 Non ☐1

4. Avez-vous un autre adulte proche de vous à qui parler de votre santé ?
(Plusieurs réponses possibles)

Votre père	☐1	*Un ami de votre famille*	☐8
Votre mère	☐2	*Votre CPE*	☐9
Votre beau-père	☐3	*Un de vos enseignants*	☐10
Votre belle-mère	☐4	*L'infirmière scolaire*	☐11
Votre frère, votre sœur	☐5	*L'assistante sociale scolaire*	☐12
Vos grands-parents	☐6	*La psychologue scolaire*	☐13
Un autre membre de votre famille	☐7	*Quelqu'un d'autre*	☐14

5. Au cours des 12 derniers mois, vous est-il arrivé… ?
(Une seule réponse possible pour chaque ligne)

	Jamais	Parfois	Souvent	Très souvent
5.1. De vous réveiller la nuit	☐1	☐2	☐3	☐4
5.2. D'avoir du mal à vous endormir	☐1	☐2	☐3	☐4
5.3. D'avoir le sentiment de ne pas être reposé après le sommeil	☐1	☐2	☐3	☐4
5.4. D'avoir le sentiment d'être décalé (s'endormir très tard, se réveiller très tard)	☐1	☐2	☐3	☐4
5.5. D'être arrivé en retard à l'école car vous ne vous étiez pas réveillé	☐1	☐2	☐3	☐4

6. D'accord/pas d'accord

Voici des phrases recueillies auprès d'adolescents, lisez chacune d'entre elles, et cochez « vrai » si elle correspond à ce que vous vivez, ou « faux » si elle ne correspond pas.

6.1. Je n'ai pas d'énergie pour l'école, pour le travail	Vrai ☐2	Faux ☐1
6.2. J'ai du mal à réfléchir	Vrai ☐2	Faux ☐1
6.3 Je sens que la tristesse, le cafard me débordent en ce moment	Vrai ☐2	Faux ☐1
6.4. Il n'y a rien qui m'intéresse, plus rien qui m'amuse	Vrai ☐2	Faux ☐1
6.5. Ce que je fais ne sert à rien	Vrai ☐2	Faux ☐1
6.6. Au fond, quand c'est comme ça, j'ai envie de mourir	Vrai ☐2	Faux ☐1
6.7. Je ne supporte pas grand-chose	Vrai ☐2	Faux ☐1
6.8. Je me sens découragé(e)	Vrai ☐2	Faux ☐1
6.9. Je dors très mal	Vrai ☐2	Faux ☐1
6.10. À l'école, au boulot, j'y arrive pas	Vrai ☐2	Faux ☐1

▶ LES QUESTIONS SUIVANTES PORTENT SUR LE TABAC

7. Au cours de votre vie, combien de fois avez-vous fumé des cigarettes ?
(Une seule réponse possible)

0 fois	1-2	3-5	6-9	10-19	20-39	40+
☐1	☐2	☐3	☐4	☐5	☐6	☐7

8. Au cours des 30 derniers jours, avez-vous fumé des cigarettes ?
(Une seule réponse possible)

Aucune	☐1	6-10 cigarettes par jour	☐5
Moins d'une cigarette par semaine	☐2	11-20 cigarettes par jour	☐6
Moins d'une cigarette par jour	☐3	Plus de 20 cigarettes par jour	☐7
1-5 cigarettes par jour	☐4		

▶ LES QUESTIONS SUIVANTES PORTENT SUR L'ALCOOL

9. Combien de fois, avez-vous bu des boissons alcoolisées ?
(Une seule réponse possible pour chaque ligne)

	0 fois	1-2	3-5	6-9	10-19	20-39	40+
9.1. Au cours de votre vie	☐1	☐2	☐3	☐4	☐5	☐6	☐7
9.2. Au cours des 12 derniers mois	☐1	☐2	☐3	☐4	☐5	☐6	☐7
9.3. Au cours des 30 derniers jours	☐1	☐2	☐3	☐4	☐5	☐6	☐7

10. Indiquez sur cette échelle de 1 à 10 à quel point vous pensez avoir été ivre le dernier jour où vous avez bu de l'alcool
(si vous n'avez ressenti aucun effet, cochez la case n° 1).

1	2	3	4	5	6	7	8	9	10
☐1	☐2	☐3	☐4	☐5	☐6	☐7	☐8	☐9	☐10

Je ne bois jamais d'alcool ☐11

11. Repensez aux 30 derniers jours, combien de fois avez-vous bu cinq « verres » ou plus en une seule occasion ? (un « verre » est un verre de vin, une canette de bière, une coupe de champagne, une bolée de cidre, un verre d'alcool fort ou un cocktail, un prémix, etc.)
(Une seule réponse possible)

Aucune	1 fois	2 fois	3-5 fois	6-9 fois	10 fois ou plus
☐1	☐2	☐3	☐4	☐5	☐6

12. Quand vous buvez de l'alcool c'est plutôt pendant :
(Plusieurs réponses possibles)

Un repas familial	☐1
Des soirées ou moments entre copains hors de l'école ou de votre lieu de travail, de stage	☐2
Des après-midi à l'école ou sur votre lieu de travail, de stage	☐3
Seul	☐4
Autre	☐5

▶ LA QUESTION SUIVANTE PORTE SUR LES MÉDICAMENTS

13. Avez-vous déjà pris des tranquillisants ou somnifères parce qu'un médecin vous a dit de les prendre ?
(Une seule réponse possible)

Non, jamais ☐1 **Oui, mais pendant moins de 3 semaines** ☐2 **Oui, pendant plus de 3 semaines** ☐3

▶ LES QUESTIONS SUIVANTES PORTENT SUR LE CANNABIS

14. Combien de fois avez-vous pris du cannabis ? (shit, joint, haschich, marijuana)
(Une seule réponse possible par ligne)

	0 fois	1-2	3-5	6-9	10-19	20-39	40+
14.1. Au cours de votre vie	☐1	☐2	☐3	☐4	☐5	☐6	☐7
14.2. Au cours des 12 derniers mois	☐1	☐2	☐3	☐4	☐5	☐6	☐7
14.3. Au cours des 30 derniers jours	☐1	☐2	☐3	☐4	☐5	☐6	☐7

15. Au cours de votre vie, vous est-il arrivé d'être dans une situation où vous auriez pu essayer de fumer du cannabis (shit, joint, haschich, marijuana) **mais vous ne l'avez pas fait ?**

Oui ☐2 Non ☐1

15.1. **Si oui, combien de fois cela vous est-il arrivé dans votre vie ?**
(Une seule réponse possible)

0 fois	1-2 fois	3-5 fois	6-9 fois	10-19 fois	20-39 fois	40 +
☐1	☐2	☐3	☐4	☐5	☐6	☐7

▶ LES QUESTIONS SUIVANTES PORTENT SUR LES AUTRES DROGUES

16. Au cours de votre vie, combien de fois avez-vous pris les drogues suivantes ?
(Une seule réponse possible par ligne)

	0 fois	1-2	3-5	6-9	10-19	20-39	40+
16.1. Tranquillisants ou Somnifères (sans ordonnance médicale)	☐1	☐2	☐3	☐4	☐5	☐6	☐7
16.2. Amphétamines	☐1	☐2	☐3	☐4	☐5	☐6	☐7
16.3. Produit à inhaler	☐1	☐2	☐3	☐4	☐5	☐6	☐7
16.4. Ecstasy	☐1	☐2	☐3	☐4	☐5	☐6	☐7
16.5. LSD ou acide	☐1	☐2	☐3	☐4	☐5	☐6	☐7
16.6. Crack	☐1	☐2	☐3	☐4	☐5	☐6	☐7
16.7. Cocaïne	☐1	☐2	☐3	☐4	☐5	☐6	☐7
16.8. Héroïne	☐1	☐2	☐3	☐4	☐5	☐6	☐7
16.9. Champignons hallucinogènes	☐1	☐2	☐3	☐4	☐5	☐6	☐7
16.10. Alcool avec du cannabis	☐1	☐2	☐3	☐4	☐5	☐6	☐7
16.11. Drogues par injection avec seringue	☐1	☐2	☐3	☐4	☐5	☐6	☐7
16.12. Alcool avec des médicaments pour ressentir certains effets	☐1	☐2	☐3	☐4	☐5	☐6	☐7
16.13. MDMA	☐1	☐2	☐3	☐4	☐5	☐6	☐7

▶ VOTRE CORPS ET VOUS

17. Quelle est votre taille ? ☐ m ☐☐ cm

18. Quel est votre poids ? ☐☐☐ kg

19. Vous pesez-vous régulièrement ? Oui ☐2 Non ☐1

Si oui :

19.1. **Chez vous** Oui ☐2 Non ☐1

19.2. **Dans un cadre médicalisé** Oui ☐2 Non ☐1

20. Pour vous, est-ce important d'avoir un corps musclé ? Oui ☐2 Non ☐1

21. Pour vous, est-ce important d'être mince ? Oui ☐2 Non ☐1

22. Avez-vous peur d'être trop gros(se) ? Oui ☐2 Non ☐1

23. Vous êtes une fille : Avez-vous déjà eu vos règles ? Oui ☐2 Non ☐1

23.1. **Si oui, la première fois à quel âge :** ☐☐ ans

24. Vous êtes un garçon : Avez-vous la voix qui a mué ?

Non	Oui	Je ne sais pas
☐1	☐2	☐3

24.1. **Si oui, à quel âge :** ☐☐ ans

25. Êtes-vous attentif à votre physique ? Oui ☐2 Non ☐1

25.1. **Si oui, vous regardez-vous régulièrement dans un miroir ?** Oui ☐2 Non ☐1

26. Quand vous pensez à votre physique ou que vous vous regardez dans un miroir ?
(Une seule réponse possible)

Vous êtes content(e)	Vous devenez triste	Vous doutez de vous	Vous vous en fichez
☐1	☐2	☐3	☐4

27. Y a-t-il des parties de votre corps qui vous déplaisent ? Oui ☐2 Non ☐1

27.1. **Si oui, lesquelles ?**
(Plusieurs réponses possibles)

Le nez	☐1	*Le ventre*	☐7	*Les fesses*	☐13
Les oreilles	☐2	*Le torse*	☐8	*Les hanches*	☐14
Les yeux	☐3	*Les seins*	☐9	*Les genoux*	☐15
Les joues	☐4	*Les bras*	☐10	*Les pieds*	☐16
Les dents	☐5	*Les cuisses*	☐11	*Autre partie*	☐17
Le visage	☐6	*Les jambes*	☐12		

28. Évitez-vous des situations où une ou plusieurs parties de votre corps seraient visibles par d'autres personnes (comme dans les vestiaires, à la piscine etc.) **?**
(Une seule réponse possible)

Jamais	Quelquefois	Souvent	Toujours
☐1	☐2	☐3	☐4

29. Diriez-vous que votre « look » est
(Plusieurs réponses possibles)

29.a **Si vous êtes un garçon**		29.b **Si vous êtes une fille**	
Viril	☐1	*Féminin*	☐1
Efféminé	☐2	*Trop masculin*	☐2
Enfantin	☐3	*Enfantin*	☐3
Adolescent	☐4	*Adolescent*	☐4
Naturel	☐5	*Naturel*	☐5
Sophistiqué	☐6	*Sophistiqué*	☐6
Athlétique	☐7	*Athlétique*	☐7
Sexy	☐8	*Sexy*	☐8
Négligé	☐9	*Négligé*	☐9
Sans particularité	☐10	*Sans particularité*	☐10

	Non	Oui
30. Attachez-vous de l'importance à la manière dont vous vous habillez?	☐1	☐2
31. Utilisez-vous des produits pour prendre soin de votre peau?	☐1	☐2
32. Avez-vous recours aux instituts de soins esthétiques?	☐1	☐2
33. Avez-vous recours à l'épilation?	☐1	☐2

33.1. **Si oui, vous vous épilez pour**
(Plusieurs réponses possibles)

Être à la mode	☐1	*Vous sentir mieux dans votre corps*	☐4
Être belle/beau	☐2	*Faire comme les autres*	☐5
Être plus séduisant(e)	☐3	*Faire du sport*	☐6

34. Vous est-il arrivé de vous faire du mal exprès?
(Une seule réponse possible)

Jamais	Rarement	Assez souvent	Très souvent
☐1	☐2	☐3	☐4

35. Avez-vous déjà participé à des « jeux » dangereux?
(Une seule réponse possible)

Jamais	Rarement	Assez souvent	Très souvent
☐1	☐2	☐3	☐4

36. Aimeriez-vous avoir ?
(Une seule réponse possible pour chaque ligne)

	Non	Éventuellement	Oui	J'en ai déjà un
36.1. Un tatouage sur le corps	☐1	☐2	☐3	☐4
36.2. Un piercing	☐1	☐2	☐3	☐4

▶ L'ALIMENTATION ET VOUS

37. Manger est pour vous
(Une seule réponse possible)

Un plaisir	Une obligation	Une contrainte, quelque chose qui vous est imposé
☐1	☐2	☐3

38. Faites-vous attention à ce que vous mangez ? Oui ☐2 Non ☐1

38.1. **Si oui, pourquoi ?**
(Plusieurs réponses possibles)

Votre poids	☐1	*Le regard des autres*	☐3
Votre santé	☐2	*Vos parents*	☐4

	Non	Oui
39. Avez-vous souffert de trop manger ?	☐1	☐2
39.1. **Si oui, est-ce toujours actuel ?**	☐1	☐2
40. Avez-vous souffert de ne pas assez manger ?	☐1	☐2
40.1. **Si oui, est-ce toujours actuel ?**	☐1	☐2
41. Y a-t-il eu des périodes de votre vie où vous étiez en surpoids ?	☐1	☐2

41.1. **Si oui, précisez**
(Plusieurs réponses possibles)

Avant l'âge de 5 ans	☐1	*Au collège*	☐3
En primaire	☐2	*Après le collège*	☐4

▶ LA SEXUALITÉ ET VOUS

42. Par qui êtes-vous attiré(e) ?
(Une seule réponse possible)

Les filles	Les garçons	Les deux	Personne
☐1	☐2	☐3	☐4

			Si oui à quel âge la 1re fois
43. Au cours de votre vie avez-vous été amoureux(se) ?	Oui ☐2	Non ☐1	☐☐ ans 43.1
44. Au cours de votre vie avez-vous dragué ?	Oui ☐2	Non ☐1	☐☐ ans 44.1

45. Au cours de votre vie avez-vous eu des rapports sexuels ? (faire l'amour)

Oui ☐2 Non ☐1

45.1. **Si oui, votre premier rapport sexuel c'était avec :** Un garçon ☐1 Une fille ☐2

45.1.1. **Quel âge aviez-vous ?** ☐☐ ans

45.1.2. **Quel âge avait votre partenaire ?** ☐☐ ans

45.1.3. **Depuis combien de temps étiez-vous ensemble ?**
(Une seule réponse possible)

Quelques années	Quelques mois	Quelques jours	Je venais de la (le) rencontrer
☐1	☐2	☐3	☐4

45.1.4. **Lors de ce premier rapport sexuel, est-ce que vous (ou votre partenaire) avez utilisé :**
(Plusieurs réponses possibles)

La pilule	☐1	*Un autre moyen de contraception*	☐3
Des préservatifs	☐2	*Aucun moyen de contraception*	☐4

45.1.5. **Étiez-vous amoureux(se) de ce partenaire ?**
(Une seule réponse possible)

Non	Oui	Je ne sais pas
☐1	☐2	☐3

46. Pour les filles : Avez-vous déjà été enceinte ? Oui ☐2 Non ☐1

46.1. **Si vous êtes une fille avez-vous déjà fait une Interruption Volontaire de Grossesse ?**

Oui ☐2 Non ☐1

46.2. **Si vous êtes un garçon votre amie, a-t-elle déjà fait une Interruption Volontaire de Grossesse ?**

Non	Oui	Je ne sais pas
☐1	☐2	☐3

▶ VOS PARENTS ET VOUS

Quel est le niveau d'études le plus élevé de : (Une seule réponse possible)	**47. Votre père**	**48. Votre mère**
N'est jamais allé à l'école	☐1	☐1
École primaire	☐2	☐2
Etudes secondaires (Brevet)	☐3	☐3
Etudes secondaires (jusqu'au bac)	☐4	☐4
Etudes supérieures (après le bac)	☐5	☐5
Diplôme professionnel (CAP-BEP)	☐6	☐6
Autre	☐7	☐7
Je ne sais pas	☐8	☐8

Quelle est la situation actuelle de vos parents ?
(Une seule réponse possible par ligne)

	En activité	Au chômage	Sans emploi par choix	Bénéficiant des minima sociaux	À la retraite	En invalidité	Je ne sais pas
49. Père	☐1	☐2	☐3	☐4	☐5	☐6	☐7
50. Mère	☐1	☐2	☐3	☐4	☐5	☐6	☐7

51. Avez-vous perdu un ou vos deux parents ? Oui ☐2 Non ☐1

51.1. **Si oui, vous aviez quel âge ?** ☐☐ ans

51.2. **Si oui, lequel ?**
(Une seule réponse possible)

Votre père	Votre mère	Les deux
☐1	☐2	☐3

51.3. **Si oui, dans quelles circonstances ?**
(Une seule réponse possible)

Accident	Maladie	Suicide	Autre	Vous ne savez pas
☐1	☐2	☐3	☐4	☐5

52. Vous vivez le plus souvent :
(Une seule réponse possible)

Avec votre père seul	☐1	En foyer	☐6
Avec votre mère seule	☐2	En famille d'accueil	☐7
Avec vos deux parents (ensemble)	☐3	En internat scolaire	☐8
Avec un parent et un beau-parent (Père-Belle Mère et/ou Mère-Beau Père)	☐4	Dans un établissement de santé	☐9
		Chez des amis	☐10
Avec vos grands-parents	☐5	Autre	☐11

53. Vos parents sont-ils divorcés ou séparés ? Oui ☐2 Non ☐1

53.1. **Si oui, depuis combien de temps ?** ☐☐ ans ☐☐ mois

54. Pour chacun de ces sujets, avec qui parlez-vous le plus facilement :
(Une seule réponse possible par ligne)

	Avec ma mère	Avec mon père	Avec un ami	Avec un frère ou une sœur	Avec personne
54.1. L'école	☐1	☐2	☐3	☐4	☐5
54.2. Sexualité	☐1	☐2	☐3	☐4	☐5
54.3. Votre famille	☐1	☐2	☐3	☐4	☐5
54.4. Sport	☐1	☐2	☐3	☐4	☐5
54.5. Votre santé	☐1	☐2	☐3	☐4	☐5
54.6. L'actualité	☐1	☐2	☐3	☐4	☐5
54.7. Internet	☐1	☐2	☐3	☐4	☐5
54.8. Vos problèmes	☐1	☐2	☐3	☐4	☐5

55. À quelle fréquence parlez-vous de :
(Une seule réponse possible par ligne)

	Rarement	Occasionnellement	Assez souvent	Très souvent
55.1. L'école	☐1	☐2	☐3	☐4
55.2. Sexualité	☐1	☐2	☐3	☐4
55.3. Votre famille	☐1	☐2	☐3	☐4
55.4. Sport	☐1	☐2	☐3	☐4
55.5. Votre santé	☐1	☐2	☐3	☐4
55.6. L'actualité	☐1	☐2	☐3	☐4
55.7. Internet	☐1	☐2	☐3	☐4
55.8. Vos problèmes	☐1	☐2	☐3	☐4

56. Diriez-vous que :

56.1. **Votre mère vous trouve belle/beau** Oui ☐2 Non ☐1

56.2. **Votre père vous trouve belle/beau** Oui ☐2 Non ☐1

57. En général, y a-t-il des disputes dans votre famille ?
(Une seule réponse possible)

Jamais	Rarement	Assez souvent	Très souvent
☐1	☐2	☐3	☐4

58. Quand ces disputes vous posent des problèmes vous pouvez en parler
(Plusieurs réponses possibles)

À vos parents	☐1	*À un autre adulte*	☐5
À votre famille (frères, sœurs, grands-parents…)	☐2	*À vos amis rencontrés uniquement sur internet*	☐6
À votre médecin	☐3	*À vos amis de la réalité (vraie vie)*	☐7
À un « psy »	☐4	*Vous n'en parlez à personne*	☐8

59. Généralement, à quel point êtes-vous satisfait(e) de :
(Une seule réponse possible par ligne)

	Très satisfait(e)	Satisfait(e)	Ni satisfait(e) ni insatisfait(e)	Pas très satisfait(e)	Pas satisfait(e) du tout
59.1. Votre relation avec votre mère	☐1	☐2	☐3	☐4	☐5
59.2. Votre relation avec votre père	☐1	☐2	☐3	☐4	☐5
59.3. Votre relation avec vos amis de la réalité (vraie vie)	☐1	☐2	☐3	☐4	☐5
59.4. Votre relation avec vos amis rencontrés uniquement sur internet	☐1	☐2	☐3	☐4	☐5

▶ VOUS ET VOUS

60. Vous considérez-vous adolescent(e) ?
(Une seule réponse possible)

Pas du tout	Plutôt non	Plutôt oui	Tout à fait
☐1	☐2	☐3	☐4

61. Pour vous, qu'est-ce qu'être adolescent(e) ?

62. Pensez-vous que l'adolescence est une période facile ?
(Une seule réponse possible)

Non	Oui	Pas toujours
☐1	☐2	☐3

63. Voici quelques phrases que vous pouvez juger vraies ou fausses :
(Une seule réponse possible)

63.1. Je souffre d'avoir un corps qui ne représente pas ce que je suis	Vrai ☐1	Faux ☐2
63.2. Avoir du muscle c'est masculin	Vrai ☐1	Faux ☐2
63.3. Être mince c'est féminin	Vrai ☐1	Faux ☐2
63.4. Le poil c'est viril	Vrai ☐1	Faux ☐2
63.5. Le poil c'est repoussant	Vrai ☐1	Faux ☐2
63.6. La beauté c'est masculin	Vrai ☐1	Faux ☐2
63.7. La beauté c'est féminin	Vrai ☐1	Faux ☐2
63.8. Pour vivre bien, il faut prendre des risques sans les calculer	Vrai ☐1	Faux ☐2
63.9. Le plus important dans la vie, c'est d'avoir le maximum de sensations	Vrai ☐1	Faux ☐2
63.10. Le plus important dans la vie, c'est d'avoir le maximum d'émotions	Vrai ☐1	Faux ☐2
63.11. Pour ne plus être angoissé, il faut réfléchir à ce qu'on ressent	Vrai ☐1	Faux ☐2
63.12. Pour ne plus être angoissé, il faut agir	Vrai ☐1	Faux ☐2
63.13. Les adultes s'inquiètent trop pour les adolescent(e)s	Vrai ☐1	Faux ☐2
63.14. Les adultes ne s'inquiètent pas assez pour les adolescent(e)s	Vrai ☐1	Faux ☐2
63.15. Les adultes posent trop de limites aux adolescent(e)s	Vrai ☐1	Faux ☐2
63.16. Les adultes font trop confiance aux adolescent(e)s	Vrai ☐1	Faux ☐2
63.17. Les adolescent(e)s ont besoin de limites pour ne pas trop se mettre en danger	Vrai ☐1	Faux ☐2
63.18. Trop de limites poussent les adolescent(e)s à prendre des risques	Vrai ☐1	Faux ☐2
63.19. Je ne supporte pas que mes parents m'interdisent des choses qu'ils font eux-mêmes (fumer, boire, etc.)	Vrai ☐1	Faux ☐2
63.20. Si j'ai des doutes sur ma vie, je me sens mal	Vrai ☐1	Faux ☐2

64. Avez-vous déjà pensé que la vie ne vaut pas la peine d'être vécue ? Oui ☐2 Non ☐1

64.1. **Si oui, que faites-vous dans ces moments-là ?**
(Plusieurs réponses possibles)

Vous vous isolez	☐1	*Vous prenez des médicaments*	☐10
Vous faites de la musique	☐2	*Vous dessinez, peignez*	☐11
Vous lisez des livres	☐3	*Vous faites du sport*	☐12
Vous jouez à des jeux vidéo	☐4	*Vous écrivez*	☐13
Vous lisez des BD	☐5	*Vous écoutez de la musique*	☐14
Vous allez voir des copains	☐6	*Vous buvez de l'alcool*	☐15
Vous recherchez des sensations	☐7	*Vous vous faites du mal*	☐16
Vous fumez du cannabis	☐8	*Rien de tout cela*	☐17
Vous fumez du tabac	☐9		

65. Au cours de votre vie, avez-vous fait une tentative de suicide ?
(Une seule réponse possible)

Non	1 fois	Plusieurs fois
☐1	☐2	☐3

66. Êtes-vous « accro » à quelque chose ? Oui ☐2 Non ☐1

66.1. **Si oui, à quoi ?**

67. Avez-vous confiance en l'avenir ?
(Une seule réponse possible)

Non	Oui	Pas toujours
☐1	☐2	☐3

68. Qu'est-ce que vous aimeriez faire qui vous est interdit à votre âge ?

69. Combien d'amis avez-vous ?
(Une seule réponse possible par ligne)

	Aucun	1	2-4	5-10	>10
69.1. Dans la réalité (vraie vie)	☐1	☐2	☐3	☐4	☐5
69.2. Rencontrés uniquement sur internet	☐1	☐2	☐3	☐4	☐5

70. Avez-vous vu en vrai des amis que vous aviez rencontré d'abord sur internet ?

Oui ☐2 Non ☐1

71. Quels sont pour vous les amis les plus importants ?
(Une seule réponse possible)

Ceux que vous avez dans la réalité (vraie vie) ☐1
Ceux que vous rencontrez uniquement sur internet ☐2

71.1. **Ce sont les plus importants car :**
(Plusieurs réponses possibles)

Vous pouvez vous confier à eux	☐1	*Vous pouvez vous prendre dans les bras*	☐4
Ils vous comprennent mieux	☐2	*Vous avez des centres d'intérêts communs*	☐5
Vous pouvez vous retrouver ensemble	☐3		

72. Allez-vous sur un réseau ? Oui ☐2 Non ☐1

72.1. **Si oui, lequel préférez-vous ?**
(Une seule réponse possible)

Facebook	Twitter	Skype
☐1	☐2	☐3

72.2. **Vous le préférez parce qu'il est ?**
(Plusieurs réponses possibles)

Confidentiel	*Vos amis y vont*	*Interactif*
☐1	☐2	☐3

73. D'après vous, que pensent de vous ?
(Une seule réponse possible par ligne)

	Du bien	Du mal	Ça m'est égal	J'aimerais le savoir	Ils ne me connaissent pas
73.1. Les autres adolescents	☐1	☐2	☐3	☐4	☐5
73.2. Vos amis	☐1	☐2	☐3	☐4	☐5
73.3. Les adultes	☐1	☐2	☐3	☐4	☐5

(Une seule réponse possible par ligne)	Jamais	Rarement	Assez souvent	Très souvent
74. Vous posez-vous des questions sur vous ?	☐1	☐2	☐3	☐4
75. Vous posez-vous des questions sur le monde qui vous entoure ?	☐1	☐2	☐3	☐4

76. Si oui, quelles questions vous posez-vous sur vous ou sur le monde qui vous entoure ?

77. Voyez-vous actuellement un(e) psychiatre ou un(e) psychologue ?

Oui ☐2 Non ☐1

77.1. **Si oui, pourquoi avez-vous consulté ?**

77.2. **Si oui, en avez-vous vu un(e) l'année dernière ?** Oui ☐2 Non ☐1

77.3. **Cela vous aide (ou vous a aidé) ?** Oui ☐2 Non ☐1

78. Conseilleriez-vous à un(e) ami(e) en difficulté d'aller en consulter un(e) ?
(Une seule réponse possible)

Oui ☐2 Non ☐1 Éventuellement ☐3

79. Pour vous, votre valeur personnelle dépend surtout

79.1. De vos résultats scolaires	Oui ☐2	Non ☐1
79.2. De votre créativité	Oui ☐2	Non ☐1
79.3. De vos performances sportives	Oui ☐2	Non ☐1
79.4. De la valeur ou du nombre des objets que vous avez	Oui ☐2	Non ☐1
79.5. De l'image que vous donnez	Oui ☐2	Non ☐1
79.6. Du nombre de vos amis	Oui ☐2	Non ☐1

80. La scolarité c'est pour vous

80.1. Enrichissant	Oui ☐2	Non ☐1
80.2. Important	Oui ☐2	Non ☐1
80.3. Stressant, énervant	Oui ☐2	Non ☐1
80.4. Inutile	Oui ☐2	Non ☐1
80.5. Agréable	Oui ☐2	Non ☐1
80.6. Obligatoire	Oui ☐2	Non ☐1
80.7. Fatigant	Oui ☐2	Non ☐1
80.8. Pénible	Oui ☐2	Non ☐1
80.9. Décourageant	Oui ☐2	Non ☐1
80.10. Utile pour l'avenir	Oui ☐2	Non ☐1
80.11. On est obligé de réussir à l'école pour réussir sa vie plus tard	Oui ☐2	Non ☐1
80.12. Peu important pour vos parents	Oui ☐2	Non ☐1
80.13. Essentiel pour vos parents	Oui ☐2	Non ☐1
80.14. La seule chose qui compte pour vos parents	Oui ☐2	Non ☐1

▶ VOS LOISIRS ET VOUS

81. Quels sont vos loisirs préférés et leur fréquence ?
(Une seule réponse possible par ligne)

	Jamais	Occasionnellement	Plusieurs fois par semaine	Chaque jour
81.1. Écouter de la musique	☐1	☐2	☐3	☐4
81.2. Faire de la musique	☐1	☐2	☐3	☐4
81.3. Être avec vos amis de la réalité	☐1	☐2	☐3	☐4
81.4. Être avec vos amis rencontrés uniquement sur internet	☐1	☐2	☐3	☐4
81.5. Lire des livres	☐1	☐2	☐3	☐4
81.6. Lire des BD	☐1	☐2	☐3	☐4
81.7. Faire du sport dans un club	☐1	☐2	☐3	☐4
81.8. Jouer à des jeux de société	☐1	☐2	☐3	☐4
81.9. Pratiquer une activité physique mais pas dans un club (roller, foot etc.)	☐1	☐2	☐3	☐4
81.10. Utiliser un ordinateur pour jouer	☐1	☐2	☐3	☐4
81.11. Utiliser un ordinateur pour aller sur des sites internet	☐1	☐2	☐3	☐4
81.12. Écrire	☐1	☐2	☐3	☐4
81.13. Dessiner, peindre	☐1	☐2	☐3	☐4

82. Faites-vous du sport régulièrement ? (Au moins une fois par semaine)

	Oui	Non
82.1. En loisir	☐2	☐1
82.2. En compétition	☐2	☐1

82.3. **Si oui, lequel ?** (celui que vous préférez pratiquer)

82.4. **Estimez-vous que c'est un sport à risque ?** Oui ☐2 Non ☐1

82.4.1. **Si oui, pensez-vous connaître vos limites ?** Oui ☐2 Non ☐1

82.5. **Combien d'heures par semaine pratiquez-vous ce sport ?**
(Une seule réponse possible)

1 heure par semaine	2 heures par semaine	3 heures par semaine	Plus de 3 heures par semaine
☐1	☐2	☐3	☐4

83. Vos motivations pour faire du sport

83.1. Être en bonne santé	Oui ☐2	Non ☐1
83.2. Vous sentir mieux dans votre corps	Oui ☐2	Non ☐1
83.3. Avoir des sensations fortes	Oui ☐2	Non ☐1
83.4. Être belle/beau	Oui ☐2	Non ☐1
83.5. Être performant(e) en compétition	Oui ☐2	Non ☐1
83.6. Prendre des risques	Oui ☐2	Non ☐1
83.7. Être en groupe	Oui ☐2	Non ☐1
83.8. Repousser vos limites	Oui ☐2	Non ☐1
83.9. Vous faire mal quand vous n'avez pas le moral	Oui ☐2	Non ☐1
83.10. Aller dans le « rouge » et ne plus penser	Oui ☐2	Non ☐1
83.11. Vous aider quand vous n'avez pas le moral	Oui ☐2	Non ☐1
83.12. Vous soigner	Oui ☐2	Non ☐1
83.13. Le plaisir	Oui ☐2	Non ☐1
83.14. Faire comme les autres	Oui ☐2	Non ☐1

84. Vous ne faites pas régulièrement du sport parce que vous
(Une seule réponse possible)

N'avez pas le temps	Avez un problème de santé	N'aimez pas le sport	Autre raison
☐1	☐2	☐3	☐4

85. Faites-vous de la musculation ? Oui ☐2 Non ☐1

Si oui, quelles sont vos motivations pour la pratique de la musculation ?

85.1. Pour le plaisir	Oui ☐2	Non ☐1
85.2. Pour vous sentir mieux dans votre corps	Oui ☐2	Non ☐1
85.3. Pour être fort(e)	Oui ☐2	Non ☐1
85.4. Pour être belle/beau	Oui ☐2	Non ☐1
85.5. Pour être plus séduisant(e)	Oui ☐2	Non ☐1
85.6. Pour être performant(e) en compétition	Oui ☐2	Non ☐1

▶ LES QUESTIONS SUIVANTES PORTENT SUR LA TÉLÉVISION (classique, téléchargement, web-tv)

86. Quels sont vos types d'émissions préférés ?
(Une seule réponse possible par ligne)

	J'adore	J'aime	Je n'aime pas trop	Je déteste
86.1. Les informations (journal de 13 h ou de 20 h par exemple)	☐1	☐2	☐3	☐4
86.2. Les magazines ou débats (politiques, littéraires…)	☐1	☐2	☐3	☐4
86.3. Les émissions culturelles, documentaires (histoire, nature…)	☐1	☐2	☐3	☐4
86.4. Les séries (comédies, fictions…)	☐1	☐2	☐3	☐4
86.5. Les films	☐1	☐2	☐3	☐4
86.6. Les divertissements (variétés, jeux)	☐1	☐2	☐3	☐4
86.7. Le sport	☐1	☐2	☐3	☐4
86.8. Les émissions de téléréalité	☐1	☐2	☐3	☐4

87. En général, à quel moment de la journée, regardez-vous la télévision ?
(Une seule réponse possible par ligne)

	Plutôt dans la journée	Plutôt le soir	Plutôt la nuit	N'importe quand	Pas du tout
87.1. En semaine	☐1	☐2	☐3	☐4	☐5
87.2. En week-end	☐1	☐2	☐3	☐4	☐5
87.3. En vacances	☐1	☐2	☐3	☐4	☐5

88. Combien de temps par jour regardez-vous la télévision ?
(Une seule réponse possible par ligne)

	Moins d'une heure	De 1 h à 3 h	Plus de 3 h
88.1. En semaine	☐1	☐2	☐3
88.2. En week-end	☐1	☐2	☐3
88.3. En vacances	☐1	☐2	☐3

89. Lorsque vous regardez la télévision à la maison c'est ?
(Une seule réponse possible par ligne)

	Dans le salon	Dans la cuisine	Dans votre chambre	Dans une autre pièce
89.1. En famille	☐1	☐2	☐3	☐4
89.2. Seul	☐1	☐2	☐3	☐4
89.3. Entre amis	☐1	☐2	☐3	☐4

90. Quels sont vos séries et vos dessins animés préférés ?
(Une seule réponse possible par ligne)

	J'adore	J'aime	Je n'aime pas trop	Je déteste	Je ne connais pas
90.1. Prison break	☐1	☐2	☐3	☐4	☐5
90.2. Malcom	☐1	☐2	☐3	☐4	☐5
90.3. Grey's anatomy	☐1	☐2	☐3	☐4	☐5
90.4. Plus belle la vie	☐1	☐2	☐3	☐4	☐5
90.5. Skins	☐1	☐2	☐3	☐4	☐5
90.6. Les Experts	☐1	☐2	☐3	☐4	☐5
90.7. Dr House	☐1	☐2	☐3	☐4	☐5
90.8. NCIS	☐1	☐2	☐3	☐4	☐5
90.9. Glee	☐1	☐2	☐3	☐4	☐5
90.10. Mentalist	☐1	☐2	☐3	☐4	☐5
90.11. Médium	☐1	☐2	☐3	☐4	☐5
90.12. Les Simpsons	☐1	☐2	☐3	☐4	☐5
90.13. Naruto	☐1	☐2	☐3	☐4	☐5
90.14. One piece	☐1	☐2	☐3	☐4	☐5

91. Quel est votre héros ou votre personnage télévisé préféré ?

91.1. **Vous le préférez car ?** Il vous ressemble ☐1 Il est différent de vous ☐2

91.2. **Ce héros est :**
(Plusieurs réponses possibles)

Drôle	☐1	*Quelqu'un de bien*	☐4	*Fort(e)*	☐7	*Solitaire*	☐9
Beau (belle)	☐2	*Violent(e)*	☐5	*Intelligent(e)*	☐8	*Peu sûr de lui (elle)*	☐10
Dangereux(se)	☐3	*Courageux(se)*	☐6				

92. Que pensez-vous de ces émissions ?
(Une seule réponse possible par ligne)

	J'adore	J'aime	Je n'aime pas trop	Je déteste	Je ne connais pas
92.1. Un dîner presque parfait	☐1	☐2	☐3	☐4	☐5
92.2. Secret story	☐1	☐2	☐3	☐4	☐5
92.3. The Voice	☐1	☐2	☐3	☐4	☐5
92.4. Tous ensemble	☐1	☐2	☐3	☐4	☐5
92.5. La belle et ses princes presque charmants	☐1	☐2	☐3	☐4	☐5
92.6. Koh lanta	☐1	☐2	☐3	☐4	☐5
92.7. Star Academy	☐1	☐2	☐3	☐4	☐5
92.8. L'amour est dans le pré	☐1	☐2	☐3	☐4	☐5
92.9. La France a un incroyable talent	☐1	☐2	☐3	☐4	☐5
92.10. L'île de la tentation	☐1	☐2	☐3	☐4	☐5
92.11. Pékin Express	☐1	☐2	☐3	☐4	☐5
92.12. L'amour est aveugle	☐1	☐2	☐3	☐4	☐5
92.13. D&Co	☐1	☐2	☐3	☐4	☐5
92.14. Les anges de la téléréalité	☐1	☐2	☐3	☐4	☐5

93. Trouvez-vous les émissions que vous préférez regarder :
(Plusieurs réponses possibles)

Intéressantes	☐1	*Choquantes pour d'autres*	☐4
Assez violentes	☐2	*Assez sexuelles*	☐5
Drôles	☐3	*Elles vous détendent*	☐6

94. Que reprochez-vous en majorité aux programmes de télévision que vous regardez ?
(Plusieurs réponses possibles)

Trop de publicité	☐1	*Trop de sexualité*	☐4
Trop violents	☐2	*Idiots, débiles*	☐5
Trop choquants	☐3		

▶ LES QUESTIONS SUIVANTES PORTENT SUR LA MUSIQUE

95. Quels sont vos goûts musicaux ?
(Une seule réponse possible par ligne)

	J'adore	J'aime	Je n'aime pas trop	Je déteste	Je ne connais pas
95.1. Rock	☐1	☐2	☐3	☐4	☐5
95.2. RAP	☐1	☐2	☐3	☐4	☐5
95.3. Soul	☐1	☐2	☐3	☐4	☐5
95.4. Métal	☐1	☐2	☐3	☐4	☐5
95.5. R'n'B'	☐1	☐2	☐3	☐4	☐5
95.6. Musique classique	☐1	☐2	☐3	☐4	☐5
95.7. Funk	☐1	☐2	☐3	☐4	☐5
95.8. Reggae	☐1	☐2	☐3	☐4	☐5
95.9. Pop	☐1	☐2	☐3	☐4	☐5
95.10. Techno	☐1	☐2	☐3	☐4	☐5
95.11. Slam	☐1	☐2	☐3	☐4	☐5
95.12. Jazz	☐1	☐2	☐3	☐4	☐5

▶ LES QUESTIONS SUIVANTES PORTENT SUR LES JEUX VIDÉO

96. Possédez-vous un ordinateur personnel ? Oui ☐2 Non ☐1

97. Jouez-vous à des jeux sur un ordinateur ? Oui ☐2 Non ☐1

97.1. **Si oui, quel âge aviez-vous quand vous avez joué avec l'ordinateur la première fois:** ☐☐ ans

98. Avez-vous une ou plusieurs consoles ?
(Plusieurs réponses possibles)

Aucune console	☐1	*Gamecube*	☐5
Playstation portable (PSP)	☐2	*Wii*	☐6
Xbox 360	☐3	*PS3*	☐7
Nintendo DS	☐4	*Autre*	☐8

99. Vous préférez jouer à des jeux
(Une seule réponse possible)

D'aventure	☐1	D'énigmes	☐6
De combat, guerre	☐2	De stratégie	☐7
De sport (foot etc.)	☐3	D'autres jeux	☐8
Ludiques (mario etc.)	☐4	Je n'ai pas de préférence	☐9
De simulation	☐5		

100. Cochez votre jeu en réseau préféré :
(Une seule réponse possible)

WOW : World of warcraft	☐1	Poker	☐7
Dofus	☐2	PES	☐8
Rappelz	☐3	Call of Duty	☐9
Counter strike	☐4	Battlefield	☐10
Starcraft 2	☐5	Je n'ai pas de préférence	☐11
Crossfire	☐6		

101. Grâce aux jeux en réseau vous est-il arrivé de rencontrer dans la réalité un partenaire de jeu, que vous ne connaissiez pas avant ?

Oui ☐2 Non ☐1

102. En général, vous jouez
(Une seule réponse possible par ligne)

	Plutôt dans la journée	Plutôt le soir	Plutôt la nuit	N'importe quand dans la journée	Pas du tout
102.1. En semaine	☐1	☐2	☐3	☐4	☐5
102.2. En week-end	☐1	☐2	☐3	☐4	☐5
102.3. En vacances	☐1	☐2	☐3	☐4	☐5

103. Combien de temps par jour jouez-vous ?
(Une seule réponse possible par ligne)

	Moins d'une heure	De 1 h à 3 h	De 4 h à 8 h	Plus de 8 h
103.1. En semaine	☐1	☐2	☐3	☐4
103.2. En week-end	☐1	☐2	☐3	☐4
103.3. En vacances	☐1	☐2	☐3	☐4

▶ LES QUESTIONS SUIVANTES PORTENT SUR LA MALADIE

104. Si vous avez une maladie ou des problèmes de santé depuis longtemps, répondez aux questions qui suivent, sinon passez à la page des commentaires

104.1. **Citez votre maladie ou vos problèmes de santé (principal)**

---

105. Est-ce un problème de santé ou une maladie familiale ? Oui ☐2 Non ☐1

105.1. **Depuis quel âge l'avez-vous ?** ☐☐ ans

106. Comment vivez-vous votre maladie ?
(Plusieurs réponses possibles)

Plutôt bien	☐1	*Une prison*	☐14
Plutôt mal	☐2	*Un secret dont je ne parle à personne*	☐15
Cela m'est indifférent	☐3	*Une particularité comme une autre*	☐16
Comme un cauchemar	☐4	*Un handicap*	☐17
Une épreuve qui me rend plus solide	☐5	*Une chose qui m'angoisse*	☐18
Une catastrophe pour mes parents	☐6	*Une chose qui me donne peur de mourir*	☐19
Cela me fait honte	☐7	*Une chose qui m'oblige à suivre des traitements que j'ai souvent envie d'arrêter pour être libre*	☐20
Une catastrophe pour mes frères et sœurs	☐8	*Un poids en plus dans la vie*	☐21
Une injustice	☐9	*Une partie de moi*	☐22
Une chose qui fait que mes parents s'occupent plus de moi	☐10	*Une chose étrangère à moi*	☐23
Une chose qui fait que le regard des autres sur moi est plutôt négatif	☐11	*Une chose qui me permet de mieux savourer les bons moments*	☐24
Une chose qui fait que le regard des autres sur moi est plutôt positif	☐12	*Un « filtre » qui me permet d'avoir de vrais amis*	☐25
Une chose qui fait que les autres ont pitié de moi	☐13	*Quelque chose de pas mal puisqu'au moins, je suis dispensé(e) de sport*	☐26

107. Prenez-vous des traitements médicamenteux pour traiter cette maladie ou ce problème de santé ?

Oui ☐2 Non ☐1

107.1. **Si oui, quels traitements prenez-vous pour traiter cette maladie ou ce problème de santé ?**
(Précisez le ou les noms)

107.2. **Prenez-vous ces traitements ?**
(Une seule réponse possible)

Tous les jours	Plusieurs fois par jour	Une fois par semaine	Une fois par mois	Autre fréquence
☐1	☐2	☐3	☐4	☐5

107.3. **Avez-vous d'autres traitements** (kinésithérapie, orthophonie etc.) **?** Oui ☐2 Non ☐1

107.3.1. **Si oui, lesquels ?** (Précisez le ou les noms)

108. Comment vivez-vous vos traitements médicamenteux ?
(Une seule réponse possible)

Plutôt bien	Plutôt mal	Cela m'est indifférent
☐1	☐2	☐3

109. Comment vivez-vous vos autres traitements ? (kinésithérapie, orthophonie etc.)
(Une seule réponse possible)

Plutôt bien	Plutôt mal	Cela m'est indifférent
☐1	☐2	☐3

110. Avez-vous l'impression de suivre correctement vos traitements médicamenteux ?

Oui ☐2 Non ☐1

110.1. **Si non, pourquoi ?**

111. Avez-vous l'impression de suivre correctement vos autres traitements ?

Oui ☐2 Non ☐1

111.1. **Si non, pourquoi ?**

112. Si vous vivez en établissement de santé, est-ce à cause de votre maladie ou de votre problème de santé ?

Oui ☐2 Non ☐1

112.1 **Si oui : cela est difficile pour vous ?**
(Une seule réponse possible)

Pas du tout	Un peu	Moyennement	Beaucoup
☐1	☐2	☐3	☐4

113. Comment vivez-vous votre maladie ou votre problème de santé ?
(Une seule réponse possible pour chaque ligne)

	Pas du tout	Un peu	Moyennement	Beaucoup
113.1. Votre maladie vous empêche d'être un adolescent comme les autres	☐1	☐2	☐3	☐4
113.2. Votre maladie vous rend plus mature que les autres	☐1	☐2	☐3	☐4
113.3. Avec votre maladie, vous gardez confiance en votre corps	☐1	☐2	☐3	☐4
113.4. Avec votre maladie, vous pensez pouvoir être aimé (d'amour) par quelqu'un	☐1	☐2	☐3	☐4
113.5. Votre corps vous paraît fragile	☐1	☐2	☐3	☐4

114. Avez-vous souvent l'impression que les adolescents, qui ne sont pas malades, sont trop superficiels ?
(Une seule réponse possible)

Beaucoup	Un peu	Pas du tout
☐1	☐2	☐3

115. Que pourrait-on faire pour vous aider à vivre votre maladie le mieux possible ?

▸ COMMENTAIRES

Si vous avez des remarques à faire sur le questionnaire vous pouvez les faire ci-dessous.

Si vous n'avez pas souhaité répondre à certaines questions, pouvez-vous nous dire pourquoi ?

Si vous souhaitez avoir plus d'informations ou discuter,
vous pouvez joindre l'infirmière, l'assistante sociale, le CPE, la vie scolaire,
le médecin scolaire de votre établissement,
ou l'éducateur de votre groupe de vie si vous êtes en internat.

Nous vous remercions d'avoir répondu à ce questionnaire avec sincérité.

Annexe 2: Auto questionnaire ADRS utilisé dans l'enquête « Processus d'Adolescence »

REPERAGE DE LA DEPRESSION DE L'ADOLESCENT – ADRS

Auto questionnaire pour les adolescents : ADRS (Adolescent Depression Rating Scale) version patient en 10 items :

Je coche « vrai » si la phrase correspond à ce que je vis, ou « faux » si elle ne correspond pas.

	VRAI	FAUX
1 – Je n'ai pas d'énergie pour l'école, pour le travail		
2 – J'ai du mal à réfléchir		
3 – je sens que la tristesse, le cafard me débordent en ce moment		
4 – Il n'y a rien qui m'intéresse, plus rien ne m'amuse		
5 – Ce que je fais ne sert à rien		
6 – Au fond, quand c'est comme ça, j'ai envie de mourir		
7 – Je ne supporte pas grand-chose		
8 – Je me sens découragé (e)		
9 – Je dors très mal		
10 – A l'école, au boulot, je n'y arrive pas		

Cotation :

Le score d'ADRS compris entre (10-10), permet l'identification d'un risque de dépression

- Modéré pour une valeur < 4 et <8
- Ou important pour un score > 8

L'ADRS est ici utilisée comme une variable qualitative, décrivant un risque de dépression au seuil >4

Référence :

Anne Revah-Levy, Boris Birmaher, Isabelle Gasquet and Bruno Falissard. The Adolescent Depression Rating Scale (ADRS) : a validation study (BMC Psychiatry 2007, 7 ;2)
Elisabeth Feur, Céline Labeyrie, Jeanne Boucher, Arianne Eïd, Sandrine Cabut, Saliha Dib, Katia Castetbon, Bruno
Falissard
Indicateurs de santé chez les collégiens et lycéens du Val-De-Marne, France en 2005 : excès pondéral, atteinte carieuse et risque de dépression (BEH , janvier 2007, 4)

Annexe 3 : Résultats de l'analyse bivariée entre la symptomatologie dépressive et ses variables explicatives selon le genre

		Filles			Garçons		
		Faible symptomatologie dépressive N=6250	Forte symptomatologie dépressive N=1263	p	Faible symptomatologie dépressive N=6331	Forte symptomatologie dépressive N=474	p
Région	rurale	2845 (45,52%)	564 (44,66%)	0,147	3017 (47,65%)	228 (48,10%)	0,282
	Montagnarde	1646 (26,34%)	365 (28,90%)		1696 (26,79%)	113 (23,84%)	
	urbaine	1759 (28,14%)	334 (26,44%)		1618 (25,56%)	133 (28,06%)	
Age en classe	<15 ans	1868 (29,89%)	339 (26,84%)	0,085	2002 (31,62%)	137 (28,90%)	0,151
	[15ans-18ans[	3696 (59,14%)	785 (62,15%)		3794 (59,93%)	286 (60,34%)	
	[18ans-20ans]	686 (10,98%)	139 (11,01%)		535 (8,45%)	51 (10,76%)	
Type établissement	Collège	2042 (33,02%)	400 (32,05%)	0,527	2326 (37,27%)	176 (37,53%)	0,951
	Lycée	4142 (66,98%)	848 (67,95%)		3915 (62,73%)	293 (62,47%)	
Etablissement	Professionnel	1763 (28,48%)	361 (28,83%)	0,825	1884 (30,15%)	128 (27,29%)	0,210
	Général	4428 (71,52%)	891 (71,17%)		4364 (69,85%)	341 (72,71%)	
Avez-vous déjà redoublé	Non	4442 (71,14%)	802 (63,60%)	<0,001	4316 (68,27%)	281 (59,41%)	<0,001
	Oui	1802 (28,86%)	459 (36,40%)		2006 (31,73%)	192 (40,59%)	
Actuellement que pensez-vous de l'école ?	J'aime	4612 (73,98%)	586 (46,40%)	<0,001	3759 (59,61%)	145 (30,66%)	<0,001
	Je n'aime pas	1622 (26,02%)	677 (53,60%)		2547 (40,39%)	328 (69,34%)	
La dernière fois que vous avez vu un médecin, c'était ?	<= 1 an	5951 (95,64%)	1209 (96,03%)	0,591	5940 (94,26%)	429 (90,89%)	0,004
	> 1 an	271 (4,36%)	50 (3,97%)		362 (5,74%)	43 (9,11%)	
Au cours des 12 derniers mois, vous est-il arrivé d'avoir le sentiment d'être décalé (s'endormir très tard et se réveiller très tard	Jamais	4563 (74,93%)	660 (53,31%)	<0,001	4722 (76,81%)	250 (54,47%)	<0,001
	Souvent	1527 (25,07%)	578 (46,69%)		1426 (23,19%)	209 (45,53%)	
Au cours des 12 derniers mois, vous est-il arrivé d'Etre en retard à l'école car vous ne vous étiez pas réveillé	Jamais	5769 (94,13%)	1071 (86,09%)	<0,001	5790 (93,25%)	381 (82,11%)	<0,001
	Souvent	360 (5,87%)	173 (13,91%)		419 (6,75%)	83 (17,89%)	
Consommation Tabac au cours des 30 derniers jours	Aucune	4304 (69,00%)	638 (50,59%)	<0,001	4421 (69,97%)	238 (50,21%)	<0,001
	< 10	1322 (21,19%)	384 (30,45%)		1194 (18,90%)	123 (25,95%)	
	>= 10	612 (9,81%)	239 (18,95%)		703 (11,13%)	113 (23,84%)	
Consommation alcool au cours des 30 derniers jours	Aucune	3487 (57,32%)	579 (47,54%)	<0,001	2883 (47,07%)	161 (36,10%)	<0,001
	< 10	1985 (32,63%)	445 (36,54%)		1919 (31,33%)	141 (31,61%)	
	>= 10	611 (10,04%)	194 (15,93%)		1323 (21,60%)	144 (32,29%)	
Intensité ivresse la dernière fois que vous avez bu	Aucune	1698 (27,85%)	223 (18,10%)	<0,001	1247 (20,19%)	84 (18,10%)	<0,001
	<5	2939 (48,20%)	539 (43,75%)		2976 (48,19%)	171 (36,85%)	
	>= 5	1460 (23,95%)	470 (38,15%)		1952 (31,61%)	209 (45,04%)	

Consommation cannabis au cours des 30 derniers jours	Aucune	5265 (85,03%)	885 (71,20%)	<0,001	4999 (79,84%)	306 (65,25%)	<0,001
	< 10	533 (8,61%)	201 (16,17%)		596 (9,52%)	72 (15,35%)	
	>= 10	394 (6,36%)	157 (12,63%)		666 (10,64%)	91 (19,40%)	
Quand vous buvez de l'alcool c'est plutôt pendant un repas familial	Non	3119 (49,90%)	695 (55,03%)	<0,001	3144 (49,66%)	282 (59,49%)	<0,001
	Oui	3131 (50,10%)	568 (44,97%)		3187 (50,34%)	192 (40,51%)	
Quand vous buvez de l'alcool c'est plutôt en soirée/copains	Non	2734 (43,74%)	412 (32,62%)	<0,001	2554 (40,34%)	161 (33,97%)	0,007
	Oui	3516 (56,26%)	851 (67,38%)		3777 (59,66%)	313 (66,03%)	
Quand vous buvez de l'alcool c'est plutôt à l'école, lieu stage/travail	Non	6181 (98,90%)	1228 (97,23%)	<0,001	6085 (96,11%)	437 (92,19%)	<0,001
	Oui	69 (1,10%)	35 (2,77%)		246 (3,89%)	37 (7,81%)	
Quand vous buvez de l'alcool c'est plutôt seul(e)	Non	6161 (98,58%)	1168 (92,48%)	<0,001	6092 (96,22%)	401 (84,60%)	<0,001
	Oui	89 (1,42%)	95 (7,52%)		239 (3,78%)	73 (15,40%)	
Quand vous buvez de l'alcool c'est plutôt durant d'Autres occasions	Non	5519 (88,30%)	1068 (84,56%)	<0,001	5223 (82,50%)	367 (77,43%)	0,007
	Oui	731 (11,70%)	195 (15,44%)		1108 (17,50%)	107 (22,57%)	
Au cours de votre vie, avez-vous consommé des amphétamines	Non	6095 (98,51%)	1192 (95,59%)	<0,001	6136 (98,46%)	436 (93,56%)	<0,001
	Oui	92 (1,49%)	55 (4,41%)		96 (1,54%)	30 (6,44%)	
Au cours de votre vie, avez-vous consommé des produits à inhaler	Non	5919 (95,71%)	1116 (89,21%)	<0,001	5954 (95,63%)	398 (85,41%)	<0,001
	Oui	265 (4,29%)	135 (10,79%)		272 (4,37%)	68 (14,59%)	
Au cours de votre vie, avez-vous consommé de l'Ecstasy	Non	6097 (98,59%)	1199 (95,77%)	<0,001	6067 (97,43%)	413 (88,44%)	<0,001
	Oui	87 (1,41%)	53 (4,23%)		160 (2,57%)	54 (11,56%)	
Au cours de votre vie, avez-vous consommé des LSD	Non	6111 (98,84%)	1209 (96,80%)	<0,001	6112 (98,00%)	433 (92,32%)	<0,001
	Oui	72 (1,16%)	40 (3,20%)		125 (2,00%)	36 (7,68%)	
Au cours de votre vie, avez-vous consommé du Crack	Non	6103 (98,72%)	1190 (95,05%)	<0,001	6089 (97,74%)	430 (92,27%)	<0,001
	Oui	79 (1,28%)	62 (4,95%)		141 (2,26%)	36 (7,73%)	
Au cours de votre vie, avez-vous consommé de la Cocaïne	Non	6024 (97,21%)	1155 (92,18%)	<0,001	6005 (96,19%)	425 (90,43%)	<0,001
	Oui	173 (2,79%)	98 (7,82%)		238 (3,81%)	45 (9,57%)	
Au cours de votre vie, avez-vous consommé de l'héroïne	Non	6129 (98,98%)	1194 (95,52%)	<0,001	6117 (98,06%)	440 (93,82%)	<0,001
	Oui	63 (1,02%)	56 (4,48%)		121 (1,94%)	29 (6,18%)	
	Non	6022 (97,33%)	1158 (92,79%)	<0,001	5924 (95,04%)	406 (86,94%)	<0,001

Au cours de votre vie, avez-vous consommé des champignons hallucinogènes	Oui	165 (2,67%)	90 (7,21%)		309 (4,96%)	61 (13,06%)	
Au cours de votre vie, avez-vous consommé de l'alcool avec du cannabis	Non	5112 (82,52%)	878 (70,13%)	<0,001	4856 (77,52%)	299 (63,62%)	<0,001
	Oui	1083 (17,48%)	374 (29,87%)		1408 (22,48%)	171 (36,38%)	
Au cours de votre vie, avez-vous consommé des drogues par injection	Non	6174 (99,76%)	1219 (97,44%)	<0,001	6175 (99,12%)	447 (95,72%)	<0,001
	Oui	15 (0,24%)	32 (2,56%)		55 (0,88%)	20 (4,28%)	
Au cours de votre vie, avez-vous consommé des alcool et médicaments	Non	6058 (97,77%)	1150 (91,63%)	<0,001	6140 (98,43%)	430 (91,88%)	<0,001
	Oui	138 (2,23%)	105 (8,37%)		98 (1,57%)	38 (8,12%)	
Au cours de votre vie, avez-vous consommé des MDMA	Non	6084 (98,72%)	1208 (97,11%)	<0,001	6129 (98,52%)	431 (93,29%)	<0,001
	Oui	79 (1,28%)	36 (2,89%)		92 (1,48%)	31 (6,71%)	
Indice de masse corporelle en 3 catégories	Sous-poids	902 (15,11%)	194 (16,06%)	<0,001	595 (9,80%)	53 (11,78%)	0,071
	Normal	4391 (73,54%)	831 (68,79%)		4660 (76,77%)	324 (72,00%)	
	Surpoids	678 (11,35%)	183 (15,15%)		815 (13,43%)	73 (16,22%)	
Vous pesez-vous régulièrement chez vous	Non	2811 (45,54%)	507 (40,59%)	0,001	3423 (54,65%)	266 (56,36%)	0,502
	Oui	3361 (54,46%)	742 (59,41%)		2841 (45,35%)	206 (43,64%)	
Vous pesez-vous régulièrement dans un cadre médicalisé	Non	3824 (66,69%)	746 (64,59%)	0,179	4133 (70,65%)	292 (66,06%)	0,048
	Oui	1910 (33,31%)	409 (35,41%)		1717 (29,35%)	150 (33,94%)	
Pour vous est-ce important d'avoir un corps musclé	Non	3286 (53,09%)	585 (46,80%)	<0,001	2159 (34,40%)	139 (29,51%)	0,035
	Oui	2904 (46,91%)	665 (53,20%)		4117 (65,60%)	332 (70,49%)	
Pour vous, est-ce important d'être mince	Non	2565 (41,35%)	344 (27,43%)	<0,001	3851 (61,44%)	252 (53,62%)	0,001
	Oui	3638 (58,65%)	910 (72,57%)		2417 (38,56%)	218 (46,38%)	
Avez-vous déjà eu vos règles/ la voix qui a mué ?	Non	394 (6,32%)	50 (3,97%)	0,002	1754 (27,93%)	114 (24,15%)	0,086
	Oui	5840 (93,68%)	1208 (96,03%)		4525 (72,07%)	358 (75,85%)	
Etes-vous attentif à votre physique ?	Non	482 (8,18%)	113 (9,45%)	0,167	1078 (17,14%)	83 (17,66%)	0,821
	Oui	5408 (91,82%)	1083 (90,55%)		5212 (82,86%)	387 (82,34%)	
Vous regardez-vous régulièrement dans un miroir	Non	1157 (19,63%)	189 (15,76%)	0,002	2336 (37,28%)	133 (28,42%)	<0,001
	Oui	4738 (80,37%)	1010 (84,24%)		3930 (62,72%)	335 (71,58%)	
Avez-vous déjà participé à des jeux dangereux	Non	6001 (96,40%)	1088 (86,49%)	<0,001	5509 (87,44%)	334 (71,22%)	<0,001
	Oui	224 (3,60%)	170 (13,51%)		791 (12,56%)	135 (28,78%)	

Manger est pour vous	Plaisir	4982 (80,56%)	830 (66,14%)	<0,001	5069 (80,79%)	319 (68,02%)	<0,001
	Contrainte	1202 (19,44%)	425 (33,86%)		1205 (19,21%)	150 (31,98%)	
Faites-vous attention à votre alimentation	Non	2134 (34,41%)	492 (39,30%)	0,001	2711 (43,28%)	267 (57,30%)	<0,001
	Oui	4067 (65,59%)	760 (60,70%)		3553 (56,72%)	199 (42,70%)	
Y-a-t-il eu des périodes de votre vie ou vous avez déjà été en surpoids	Non	4744 (76,84%)	848 (68,06%)	<0,001	4984 (79,65%)	325 (70,35%)	<0,001
	Oui	1430 (23,16%)	398 (31,94%)		1273 (20,35%)	137 (29,65%)	
Attirance sexuelle	Aucune	121 (1,94%)	17 (1,35%)	<0,001	93 (1,48%)	7 (1,48%)	<0,001
	Hétérosexuelle	5786 (92,98%)	1108 (88,01%)		6067 (96,24%)	435 (91,97%)	
	LGB	316 (5,08%)	134 (10,64%)		144 (2,28%)	31 (6,55%)	
Avez-vous déjà été amoureux(se)	Non	826 (13,32%)	121 (9,60%)	<0,001	674 (10,74%)	37 (7,84%)	0,057
	Oui	5377 (86,68%)	1139 (90,40%)		5601 (89,26%)	435 (92,16%)	
Avez-vous déjà dragué	Non	2923 (47,48%)	516 (41,28%)	<0,001	1416 (22,82%)	92 (19,83%)	0,153
	Oui	3233 (52,52%)	734 (58,72%)		4789 (77,18%)	372 (80,17%)	
Au cours de votre vie, avez-vous déjà eu des rapports sexuels ?	Non	4191 (67,35%)	727 (57,84%)	<0,001	3756 (59,76%)	194 (41,19%)	<0,001
	Oui	2032 (32,65%)	530 (42,16%)		2529 (40,24%)	277 (58,81%)	
1er rapport sexuel avec un inconnu	Non	6079 (98,11%)	1194 (95,60%)	<0,001	5801 (92,56%)	405 (85,99%)	<0,001
	Oui	117 (1,89%)	55 (4,40%)		466 (7,44%)	66 (14,01%)	
1er rapport sexuel non protégé	Non	6043 (96,69%)	1170 (92,64%)	<0,001	6104 (96,41%)	435 (91,77%)	<0,001
	Oui	207 (3,31%)	93 (7,36%)		227 (3,59%)	39 (8,23%)	
Interruption volontaire de grossesse (IVG):	Non	5673 (98,00%)	1129 (95,11%)	<0,001	5363 (97,33%)	344 (92,23%)	<0,001
	Oui	116 (2,00%)	58 (4,89%)		147 (2,67%)	29 (7,77%)	
Niveau d'étude le plus élevé du père	<BAC	2591 (54,99%)	561 (61,58%)	<0,001	2612 (51,94%)	205 (54,96%)	0,283
	>=BAC	2121 (45,01%)	350 (38,42%)		2417 (48,06%)	168 (45,04%)	
Niveau d'étude le plus élevé de la mère	<BAC	2297 (45,23%)	499 (51,34%)	0,001	1989 (38,74%)	156 (41,05%)	0,403
	>=BAC	2781 (54,77%)	473 (48,66%)		3145 (61,26%)	224 (58,95%)	
Catégorie socioprofessionnelle du père	En activité	443 (7,09%)	100 (7,92%)	0,008	389 (6,14%)	45 (9,49%)	<0,001
	Chômage	5256 (84,10%)	1020 (80,76%)		5517 (87,14%)	383 (80,80%)	
	NSP	551 (8,82%)	143 (11,32%)		425 (6,71%)	46 (9,70%)	
Catégorie socioprofessionn-elle de la mère	En activité	970 (15,52%)	239 (18,92%)	<0,001	785 (12,40%)	84 (17,72%)	0,001

	Chômage	5015 (80,24%)	948 (75,06%)		5309 (83,86%)	367 (77,43%)	
	NSP	265 (4,24%)	76 (6,02%)		237 (3,74%)	23 (4,85%)	
Avez-vous perdu au moins un parent	Non	5846 (95,30%)	1159 (93,92%)	0,048	5823 (95,46%)	428 (93,45%)	0,065
	Oui	288 (4,70%)	75 (6,08%)		277 (4,54%)	30 (6,55%)	
Suicide d'au moins un parent	Non	6101 (99,56%)	1217 (98,62%)	<0,001	6064 (99,46%)	450 (98,25%)	0,007
	Oui	27 (0,44%)	17 (1,38%)		33 (0,54%)	8 (1,75%)	
Vivez-vous avec au moins un parent	Non	601 (9,76%)	184 (14,78%)	<0,001	608 (9,82%)	74 (15,95%)	<0,001
	Oui	5555 (90,24%)	1061 (85,22%)		5584 (90,18%)	390 (84,05%)	
Séparation/divorce des parents	Non	3966 (65,66%)	713 (58,35%)	<0,001	4081 (67,97%)	253 (56,85%)	<0,001
	Oui	2074 (34,34%)	509 (41,65%)		1923 (32,03%)	192 (43,15%)	
Parlez-vous le plus facilement avec vos parents de l'école	Non	1464 (26,99%)	446 (40,07%)	<0,001	1374 (25,89%)	195 (47,79%)	<0,001
	Oui	3960 (73,01%)	667 (59,93%)		3933 (74,11%)	213 (52,21%)	
Parlez-vous le plus facilement avec vos parents de sexualité	Non	4011 (71,21%)	924 (79,59%)	<0,001	4471 (80,17%)	345 (83,54%)	0,110
	Oui	1622 (28,79%)	237 (20,41%)		1106 (19,83%)	68 (16,46%)	
Parlez-vous le plus facilement avec vos parents de votre famille	Non	2771 (51,75%)	736 (65,89%)	<0,001	2251 (44,67%)	235 (59,19%)	<0,001
	Oui	2584 (48,25%)	381 (34,11%)		2788 (55,33%)	162 (40,81%)	
Parlez-vous le plus facilement avec vos parents de sport	Non	2689 (49,97%)	622 (56,60%)	<0,001	2394 (46,25%)	234 (59,09%)	<0,001
	Oui	2692 (50,03%)	477 (43,40%)		2782 (53,75%)	162 (40,91%)	
Parlez-vous le plus facilement avec vos parents de votre santé	Non	893 (16,60%)	386 (34,74%)	<0,001	925 (18,17%)	155 (39,34%)	<0,001
	Oui	4486 (83,40%)	725 (65,26%)		4167 (81,83%)	239 (60,66%)	
Parlez-vous le plus facilement avec vos parents de l'actualité	Non	2472 (45,86%)	622 (55,73%)	<0,001	2119 (41,03%)	206 (52,15%)	<0,001
	Oui	2918 (54,14%)	494 (44,27%)		3045 (58,97%)	189 (47,85%)	
Parlez-vous le plus facilement avec vos parents d'internet	Non	3943 (72,07%)	868 (78,06%)	<0,001	3676 (69,49%)	318 (77,37%)	0,001
	Oui	1528 (27,93%)	244 (21,94%)		1614 (30,51%)	93 (22,63%)	
Parlez-vous le plus facilement avec vos parents de vos problèmes	Non	3630 (65,10%)	932 (81,40%)	<0,001	3158 (59,84%)	311 (74,94%)	<0,001
	Oui	1946 (34,90%)	213 (18,60%)		2119 (40,16%)	104 (25,06%)	
En général, y-a-t-il des disputes dans votre famille	Jamais	3688 (59,29%)	413 (32,88%)	<0,001	4603 (73,20%)	220 (47,21%)	<0,001
	Souvent	2532 (40,71%)	843 (67,12%)		1685 (26,80%)	246 (52,79%)	

Combien d'amis avez-vous dans la réalité	Aucun	85 (1,37%)	33 (2,62%)	<0,001	83 (1,32%)	18 (3,82%)	<0,001
	<10	2342 (37,68%)	538 (42,66%)		1431 (22,77%)	121 (25,69%)	
	>=10	3788 (60,95%)	690 (54,72%)		4771 (75,91%)	332 (70,49%)	
Combien d'amis avez-vous rencontrés uniquement sur internet	Aucun	3568 (59,50%)	603 (49,22%)	<0,001	2722 (44,62%)	188 (40,43%)	0,014
	<10	1533 (25,56%)	377 (30,78%)		1530 (25,08%)	106 (22,80%)	
	>=10	896 (14,94%)	245 (20,00%)		1848 (30,30%)	171 (36,77%)	
Vos amis pensent du bien de vous	Non	1266 (20,47%)	421 (33,73%)	<0,001	1365 (21,93%)	178 (38,36%)	<0,001
	Oui	4919 (79,53%)	827 (66,27%)		4860 (78,07%)	286 (61,64%)	
La scolarité c'est pour vous la seule chose qui compte pour vos parents	Non	4491 (75,61%)	681 (56,33%)	<0,001	4306 (73,07%)	243 (54,61%)	<0,001
	Oui	1449 (24,39%)	528 (43,67%)		1587 (26,93%)	202 (45,39%)	
La scolarité c'est pour vous peu important pour vos parents	Non	5605 (94,79%)	1114 (92,60%)	0,003	5501 (93,65%)	394 (88,74%)	<0,001
	Oui	308 (5,21%)	89 (7,40%)		373 (6,35%)	50 (11,26%)	
La scolarité c'est pour vous essentiel pour vos parents	Non	754 (12,48%)	152 (12,43%)	0,996	730 (12,18%)	60 (13,25%)	0,554
	Oui	5286 (87,52%)	1071 (87,57%)		5263 (87,82%)	393 (86,75%)	
Fréquence loisir préféré : Ecouter de la musique	Jamais	339 (5,48%)	47 (3,75%)	0,015	821 (13,22%)	45 (9,72%)	0,036
	Souvent	5850 (94,52%)	1206 (96,25%)		5387 (86,78%)	418 (90,28%)	
Fréquence loisir préféré : Faire de la musique	Jamais	5160 (84,74%)	1018 (82,83%)	0,101	5151 (85,21%)	359 (79,60%)	0,002
	Souvent	929 (15,26%)	211 (17,17%)		894 (14,79%)	92 (20,40%)	
Fréquence loisir préféré : Être avec vos amis de la réalité	Jamais	733 (11,97%)	204 (16,40%)	<0,001	806 (13,13%)	81 (17,65%)	0,008
	Souvent	5391 (88,03%)	1040 (83,60%)		5331 (86,87%)	378 (82,35%)	
Fréquence loisir préféré : Être avec vos amis internet	Jamais	5595 (92,28%)	1068 (87,18%)	<0,001	5049 (83,62%)	366 (80,62%)	0,111
	Souvent	468 (7,72%)	157 (12,82%)		989 (16,38%)	88 (19,38%)	
Fréquence loisir préféré : Lire livres/BD	Jamais	4289 (69,99%)	926 (74,86%)	0,001	4586 (75,27%)	357 (78,81%)	0,102
	Souvent	1839 (30,01%)	311 (25,14%)		1507 (24,73%)	96 (21,19%)	
Fréquence loisir préféré : Faire du sport en club/loisir	Jamais	2546 (41,54%)	568 (45,62%)	0,009	1324 (21,49%)	146 (31,53%)	<0,001
	Souvent	3583 (58,46%)	677 (54,38%)		4838 (78,51%)	317 (68,47%)	

Fréquence loisir préféré : Jouer à des jeux de société	Jamais	5871 (95,99%)	1196 (96,69%)	0,286	5712 (93,85%)	430 (94,09%)	0,918
	Souvent	245 (4,01%)	41 (3,31%)		374 (6,15%)	27 (5,91%)	
Fréquence loisir préféré : Utiliser ordinateur pour jouer	Jamais	3638 (59,44%)	681 (55,41%)	0,010	2602 (42,45%)	182 (39,48%)	0,232
	Souvent	2482 (40,56%)	548 (44,59%)		3528 (57,55%)	279 (60,52%)	
Fréquence loisir préféré : utiliser ordinateur pour aller sur internet	Jamais	1314 (21,55%)	222 (17,95%)	0,005	1445 (23,61%)	87 (19,12%)	0,033
	Souvent	4783 (78,45%)	1015 (82,05%)		4676 (76,39%)	368 (80,88%)	
Fréquence loisir préféré : Ecrire	Jamais	4049 (66,15%)	766 (61,92%)	0,005	4667 (76,66%)	355 (78,89%)	0,306
	Souvent	2072 (33,85%)	471 (38,08%)		1421 (23,34%)	95 (21,11%)	
Fréquence loisir préféré : Dessiner, peindre	Jamais	4896 (79,57%)	945 (76,27%)	0,010	5307 (86,55%)	386 (84,84%)	0,339
	Souvent	1257 (20,43%)	294 (23,73%)		825 (13,45%)	69 (15,16%)	
Pratiquez-vous un sport régulièrement	Non	1504 (24,28%)	396 (31,55%)	<0,001	619 (9,95%)	70 (15,02%)	0,002
	Loisir	3076 (49,66%)	537 (42,79%)		2506 (40,30%)	182 (39,06%)	
	Compétition	1614 (26,06%)	322 (25,66%)		3093 (49,74%)	214 (45,92%)	
Pratiquez-vous un sport à risque	Non	4327 (72,20%)	845 (69,49%)	0,060	3545 (58,44%)	247 (54,65%)	0,126
	Oui	1666 (27,80%)	371 (30,51%)		2521 (41,56%)	205 (45,35%)	
Pensez-vous Connaitre vos limites dans la pratique d'un sport à risque	Non	4393 (71,88%)	930 (75,24%)	0,017	3747 (61,46%)	294 (64,47%)	0,219
	Oui	1719 (28,12%)	306 (24,76%)		2350 (38,54%)	162 (35,53%)	
Faites-vous de la musculation ?	Non	4714 (78,27%)	934 (76,56%)	0,202	3000 (49,62%)	195 (43,43%)	0,013
	Oui	1309 (21,73%)	286 (23,44%)		3046 (50,38%)	254 (56,57%)	
Possédez-vous un ordinateur personnel	Non	1649 (27,51%)	310 (25,41%)	0,142	1724 (29,13%)	113 (25,62%)	0,130
	Oui	4345 (72,49%)	910 (74,59%)		4194 (70,87%)	328 (74,38%)	
Jouez-vous à des jeux sur un sur un ordinateur	Non	2740 (45,77%)	566 (46,47%)	0,676	1620 (27,43%)	120 (27,40%)	1,000
	Oui	3247 (54,23%)	652 (53,53%)		4287 (72,57%)	318 (72,60%)	
Temps de jeux vidéo par jour en semaine	<1H	3996 (82,34%)	784 (77,47%)	<0,001	2795 (51,59%)	204 (49,64%)	<0,001
	[1H-3H[	714 (14,71%)	176 (17,39%)		1912 (35,29%)	111 (27,01%)	
	>3H	143 (2,95%)	52 (5,14%)		711 (13,12%)	96 (23,36%)	
Temps de jeux vidéo par jour le Week end	<1H	2659 (52,87%)	511 (48,81%)	<0,001	993 (17,81%)	71 (16,86%)	<0,001
	[1H-3H[	1835 (36,49%)	362 (34,57%)		2430 (43,60%)	145 (34,44%)	

	>3H	535 (10,64%)	174 (16,62%)		2151 (38,59%)	205 (48,69%)	
Temps de jeux vidéo par jour pendant les vacances	<1H	2082 (40,77%)	397 (37,59%)	<0,001	828 (14,85%)	51 (12,14%)	<0,001
	[1H-3H[	1895 (37,11%)	349 (33,05%)		1680 (30,13%)	97 (23,10%)	
	>3H	1130 (22,13%)	310 (29,36%)		3068 (55,02%)	272 (64,76%)	

Annexe 4: Distribution du poids latent de sélection des variables explicatives selon le genre par ordre décroissant

Garçons		Filles	
Variable explicative	Poids de sélection	Variable explicative	Poids de sélection
scolarisation au collège ou lycée	**0,9998**	**Au cours de votre vie, avez-vous déjà consommé de l'alcool avec du cannabis**	**1,0000**
scolarisation en enseignement général	**0,9998**	**scolarisation en enseignement général**	**1,0000**
Intensité de l'ivresse la dernière fois que vous avez bu	**0,9995**	**Age en 3 classes**	**1,0000**
scolarisation en enseignement professionnel	**0,9994**	**scolarisation en enseignement professionnel**	**0,9998**
consommation intensive cannabis au cours des 30 derniers jours	**0,9994**	**scolarisation au collège ou lycée**	**0,9996**
consommation intensive de tabac au cours des 30 derniers jours	**0,9991**	**consommation intensive de cannabis au cours des 30 derniers jours**	**0,9994**
Age en 3 classes	**0,9989**	**Intensité de l'ivresse la dernière fois que vous avez bu**	**0,9993**
scolarisation en enseignement agricole	**0,9988**	**consommation intensive de tabac au cours des 30 derniers jours**	**0,9991**
Au cours de votre vie, avez-vous déjà consommé de l'alcool avec du cannabis ?	**0,9986**	**temps de jeux vidéo par jour en week-end**	**0,9990**
consommation intensive d'alcool au cours des 30 derniers jours ?	**0,9983**	**Pratiquez-vous régulièrement un sport ?**	**0,9990**
Quand vous buvez c'est plutôt en soirée ou avec des copains	**0,9981**	**Quand vous buvez c'est plutôt en soirée ou avec des copains**	**0,9989**
temps de jeux vidéo par jour en week-end	**0,9979**	**consommation intensive d'alcool durant les 30 derniers jours**	**0,9988**
Au cours de votre vie, avez-vous déjà consommé des champignons hallucinogènes ?	**0,9978**	**Pratiquez-vous un sport à risque ?**	**0,9981**
Au cours de votre vie, avez-vous déjà consommé de l'ecstasy	**0,9977**	**temps de jeux vidéo par jour en semaine**	**0,9981**
Au cours de votre vie, avez-vous déjà consommé de l'héroïne	**0,9970**	**Au cours de votre vie, avez-vous déjà consommé des champignons hallucinogènes**	**0,9977**
temps de jeux vidéo par jour en semaine	**0,9964**	**Avez-vous déjà eu vos règles ?**	**0,9974**
Au cours de votre vie, avez-vous déjà consommé du crack ?	**0,9959**	**scolarisation en enseignement agricole**	**0,9967**
Au cours de votre vie, avez-vous déjà consommé de la cocaïne ?	**0,9959**	**Au cours de votre vie, avez-vous déjà eu des rapports sexuels ?**	**0,9966**
Au cours de votre vie, avez-vous déjà consommé du LSD ?	**0,9950**	**Au cours de votre vie, avez-vous déjà consommé de la cocaïne ?**	**0,9963**
Au cours de votre vie, avez-vous déjà consommé de la MDMA ?	**0,9947**	**Au cours de votre vie, avez-vous déjà consommé de l'ecstasy ?**	**0,9958**
1er rapport sexuel avec quelqu'un que vous venez de rencontrer	**0,9946**	**Au cours de votre vie, avez-vous déjà consommé de l'alcool avec médicaments pour ressentir des effets**	**0,9955**
Avez-vous déjà eu des rapports sexuels ?	**0,9907**	**Au cours de votre vie, avez-vous déjà consommé de la MDMA ?**	**0,9954**
Au cours de votre vie, avez-vous déjà consommé des drogues par injection avec seringues	**0,9900**	**Au cours de votre vie, avez-vous déjà consommé du LSD ?**	**0,9951**

Garçons		Filles	
Variable explicative	Poids de sélection	Variable explicative	Poids de sélection
Au cours de votre vie, avez-vous déjà consommé de l'alcool avec médicaments pour ressentir des effets ?	**0,9879**	**Au cours de votre vie, avez-vous déjà consommé de l'héroïne ?**	**0,9948**
temps de jeux vidéo par jour en vacances	**0,9878**	**Au cours de votre vie, avez-vous consommé du crack ?**	**0,9916**
Interruption volontaire de grossesse de votre petite amie	**0,9807**	**temps de jeux vidéo par jour en vacances**	**0,9904**
Au cours de votre vie, avez-vous déjà consommé des amphétamines ?	**0,9764**	**Avez-vous déjà redoublé ?**	**0,9903**
boire à l'école ou lieu de stage ou travail	**0,9750**	**connaitre ses limites dans la pratique d'un sport à risque**	**0,9891**
Parlez-vous le plus facilement avec vos parents de l'actualité ?	**0,9723**	**Au cours de votre vie, avez-vous déjà consommé des produits à inhaler ?**	**0,9879**
attirance sexuelle	**0,9661**	**Rapport sexuel précoce dans la relation**	**0,9871**
1er rapport sexuel non protégé ?	**0,9557**	**1er rapport sexuel non protégé ?**	**0,9850**
Fréquence loisir préféré: être avec amis rencontrés sur internet	0,9431	**Interruption volontaire de grossesse**	**0,9841**
Parlez-vous le plus facilement avec vos parents de sport	0,9401	**Quand vous buvez de l'alcool c'est plutôt seul ?**	**0,9825**
Avez-vous déjà dragué ?	0,9400	**Participez-vous à des jeux dangereux ?**	**0,9744**
Au cours de votre vie, avez-vous déjà consommé des produits à inhaler	0,9302	**attirance sexuelle**	**0,9570**
Fréquence loisir préféré: être avec amis de la réalité ?	0,9145	Fréquence loisir préféré: faire du sport (en club ou non)	0,9394
Fréquence loisir préféré: faire du sport (en club ou non)	0,8978	Au cours de votre vie, avez-vous consommé des amphétamines ?	0,9335
Pratiquez-vous régulièrement un sport ?	0,8933	Fréquence loisir préféré: être avec amis de la réalité	0,8968
Participez-vous à des jeux dangereux ?	0,8926	Parlez-vous le plus facilement avec vos parents de ses problèmes ?	0,8927
Parlez-vous le plus facilement avec vos parents d'internet	0,8871	Avez-vous déjà dragué ?	0,8466
Parlez-vous facilement avec vos parents de votre famille ?	0,8843	Etes-vous attentif à votre physique ?	0,8411
Parlez-vous facilement avec vos parents de vos problèmes	0,8810	Avez-vous déjà été amoureux ?	0,8405
Quand vous buvez de l'alcool c'est plutôt seul	0,8761	faites-vous de la musculation ?	0,8289
Avez-vous déjà redoublé ?	0,8735	Parlez-vous le plus facilement avec vos parents d'internet ?	0,8259
Avez-vous déjà été amoureux ?	0,8478	Jouez-vous à des jeux sur un ordinateur ?	0,8207
Fréquence loisir préféré: utiliser un ordinateur pour jouer	0,8476	être arrivé en retard à l'école car pas réveillé	0,8050
Parlez-vous facilement avec vos parents de l'école ?	0,8471	Parlez-vous facilement avec vos parents de sexualité ?	0,7918
Fréquence loisir préféré: utiliser un ordinateur pour aller sur internet	0,8367	Généralement, il y a-t-il des disputes dans votre famille ?	0,7479
Fréquence loisir préféré: écouter de la musique	0,8329	Fréquence loisir préféré: lire des livres/BD	0,7455

Garçons		Filles	
Variable explicative	Poids de sélection	Variable explicative	Poids de sélection
être arrivé en retard à l'école car pas réveillé	0,8198	vivez-vous avec au moins un de vos parents ?	0,7438
Etes-vous attentif à votre physique ?	0,8124	Fréquence loisir préféré: utiliser un ordinateur pour aller sur internet	0,7426
Avez-vous déjà mué ?	0,7975	Parlez-vous facilement avec vos parents de sport ?	0,7378
faites-vous de la musculation ?	0,7794	Fréquence loisir préféré: faire de la musique	0,7345
Pour vous, est-ce important d'avoir un corps musclé ?	0,7554	Pour vous, est-ce important d'avoir un corps musclé ?	0,7322
sentiment d'être décalé (s'endormir trop tard)	0,7546	Parlez-vous facilement avec vos parents de votre santé ?	0,6918
Jouez-vous à des jeux sur ordinateur	0,7373	Actuellement, que pensez-vous de l'école ?	0,6823
Parlez-vous facilement avec vos parents sexualité ?	0,7355	sentiment d'être décalé (s'endormir trop tard)	0,6812
connaitre ses limites dans la pratique d'un sport à risque	0,6989	Parlez-vous facilement avec vos parents de votre famille ?	0,6808
Quand vous buvez de l'alcool c'est plutôt durant un repas familial ?	0,6802	Fréquence loisir préféré: utiliser un ordinateur pour jouer	0,6799
Vous regardez vous régulièrement dans un miroir ?	0,6521	Quand vous buvez de l'alcool c'est plutôt durant un repas familial ?	0,6664
Possédez-vous un ordinateur personnel ?	0,6484	Fréquence loisir préféré: être avec amis rencontrés sur internet	0,6526
Fréquence loisir préféré: lire des livres/BD	0,6247	Quand vous buvez de l'alcool c'est plutôt durant une autre occasion	0,6478
Faites-vous attention à ce que vous mangez ?	0,6233	Combien d'amis avez-vous rencontré uniquement sur internet ?	0,6452
Actuellement, que pensez-vous de l'école	0,6070	vous regardez régulièrement dans un miroir	0,6267
Vivez-vous avec au moins un de vos parents ?	0,6063	Parlez-vous facilement avec vos parents de l'actualité ?	0,6218
Combien d'amis avez-vous rencontrés uniquement sur internet ?	0,6012	Fréquence loisir préféré: écouter de la musique	0,6192
pratique d'un sport à risque	0,5987	Selon vous la scolarité c'est pour vous: la seule chose qui compte pour vos parents	0,6181
Selon vous, la scolarité c'est: la seule chose qui compte pour vos parents	0,5436	niveau d'études le plus élevé de la mère	0,6042
Généralement, il y a-t-il des disputes dans votre famille ?	0,5282	Fréquence Loisir préféré: dessiner ou peindre	0,5870
Vous pesez régulièrement à domicile ?	0,5155	niveau d'études le plus élevé du père	0,5748
Pour vous, est-ce important d'être mince ?	0,5045	Avez-vous déjà été en Période de surpoids dans la vie	0,5694
IMC	0,4970	Possédez-vous un ordinateur personnel ?	0,5529

Garçons		Filles	
Variable explicative	Poids de sélection	Variable explicative	Poids de sélection
Selon vous la scolarité est : peu important pour vos parents	0,4774	Pour vous est-ce important d'être mince ?	0,5461
nombre d'amis sur internet	0,4771	IMC	0,5396
Quand vous buvez de l'alcool c'est plutôt durant une autre occasion	0,4713	Catégorie socio professionnelle de la mère	0,5318
Fréquence loisir préféré: faire de la musique	0,4458	Parlez-vous facilement avec vos parents de l'école ?	0,5313
Fréquence Loisir préféré: dessiner ou peindre	0,4231	Faites-vous attention à ce que vous mangez ?	0,5284
niveau d'études le plus élevé du père	0,4191	Selon vous la scolarité est: peu importante pour vos parents	0,5155
niveau d'études le plus élevé de la mère	0,4147	Manger est pour vous un plaisir ou une contrainte	0,5131
Manger est pour vous un plaisir ou une contrainte ?	0,4008	Combien d'amis avez-vous dans la réalité	0,4897
Vous pesez-vous régulièrement dans un cadre médicalisée ?	0,3944	Catégorie socio professionnelle du père	0,4690
Vos parents sont-ils séparés ou divorcés ?	0,3917	Vos parents sont-ils séparés ou divorcés ?	0,4168
Combien d'amis avez-vous rencontrés uniquement sur internet ?	0,3916	Vous pesez-vous régulièrement à votre domicile ?	0,3943
Fréquence loisir préféré: jouer à des jeux de sociétés	0,2805	Combien d'amis avez-vous rencontrés uniquement sur internet	0,3503
Parlez-vous facilement avec vos parents de votre santé	0,2733	Vous pesez-vous régulièrement dans un cadre médicalisée ?	0,3094
Catégorie socio professionnelle du père	0,2552	Selon vous la scolarité est : essentielle pour vos parents	0,3078
y-a-t-il des périodes de votre vie ou vous avez déjà été en surpoids ?	0,2424	Fréquence loisir préféré: écrire	0,3034
Selon vous la scolarité est: essentielle pour vos parents	0,1536	Avez-vous perdu au moins un parent ?	0,0564
Avez-vous perdu au moins un parent ?	0,1179	Au cours de votre vie, avez-vous déjà consommé des drogues par injection avec seringues	0,0349
La dernière fois que vous avez-vu un médecin c'était ?	0,0461	Suicide d'au moins un parent	0,0000
Suicide d'au moins un parent	0,0000	Quand vous buvez de l'alcool c'est plutôt à l'école ou lieu de stage ou travail	0,0000
Catégorie socio professionnelle de la mère	0,0000	La dernière fois que vous avez-vu un médecin c'était ?	0,0000
Fréquence loisir préféré: écrire	0,0000	Fréquence loisir préféré: jouer à des jeux de sociétés	0,0000

Légende : **En gras**, les variables explicatives ayant un poids de sélection latent supérieur ou égal à 0,9

Bibliographie

1. Choquet M, Ledoux S. Adolescents. Enquête Nationale. 1993.

2. Jousselme C, Cosquer M, Hassler C. Portraits d'adolescents. Enquête épidémiologique multicentrique en milieu scolaire en 2013. 2015. 182 p.

3. Savalle C. Le mal-être des adolescents et sa prise en charge en santé scolaire. 2014.

4. Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE. Lifetime Prevalence and Age-of-Onset Distributions of DSM-IV Disorders in the National Comorbidity Survey Replication. Arch Gen Psychiatry. 1 juin 2005;62(6):593.

5. Hickie IB. Youth mental health: we know where we are and we can now say where we need to go next: Developing youth mental health services. Early Interv Psychiatry. févr 2011;5:63-9.

6. Colman I, Murray J, Abbott RA, Maughan B, Kuh D, Croudace TJ, et al. Outcomes of conduct problems in adolescence: 40 year follow-up of national cohort. BMJ. 8 janv 2009;338(jan08 2):a2981-a2981.

7. Dykxhoorn J, Hatcher S, Roy-Gagnon M-H, Colman I. Early life predictors of adolescent suicidal thoughts and adverse outcomes in two population-based cohort studies. Abe T, éditeur. PLOS ONE. 10 août 2017;12(8):e0183182.

8. Patton GC, Olsson C, Bond L, Toumbourou JW, Carlin JB, Hemphill SA, et al. Predicting female depression across puberty: a two-nation longitudinal study. J Am Acad Child Adolesc Psychiatry. 2008;47(12):1424-32.

9. Catry C, Braconnier A, Marcelli D. Dépressions à l'adolescence. EMC - Psychiatr. janv 2007;4(4):1-9.

10. Haute Autorité de Santé. La dépression de l'adolescent comment repérer et prendre en charge ? 2014 p. 7.

11. Sznajder M, Speranza M, Guyot C, Martin S, Nathanson S, Kerbourc'h S, et al. Depressive symptoms among teenagers in the emergency department: prevalence estimate and concordance with parental perceptions. Eur J Pediatr. déc 2013;172(12):1587-96.

12. Brunelle J, Cohen D. La dépression chez l'adolescent [Internet]. Fondation Pierre Deniker; 2015 p. 1-21. (Livre blanc de la dépression). Disponible sur: http://www.fondationpierredeniker.org/uploads/factSheets/d212cf94a002444917079a20a52acc06323188c9.pdf

13. DSM-V. Diagnostic and Statistical Manual of Mental Disorders [Internet]. Fifth Edition. American Psychiatric Association; 2013 [cité 12 déc 2020]. Disponible sur: http://psychiatryonline.org/doi/book/10.1176/appi.books.9780890425596

14. Choquet M. Suicide et adolescence : acquis épidémiologiques [Internet]. Paris : Inserm; 2000 [cité 10 déc 2020]. Disponible sur: http://www.psydoc-france.fr/conf&rm/conf/confsuicide/choquet.html

15. Hawton K, van Heeringen K. Suicide. Lancet. 2009;373.

16. Moro M-R, Brison J-L. Mission Bien Être et Santé Des Jeunes. 2016;198.

17. Santé M des S et de la, Santé M des S et de la. La stratégie nationale de santé 2018-2022 [Internet]. 2020 déc [cité 10 déc 2020]. Disponible sur: https://solidarites-sante.gouv.fr/systeme-de-sante-et-medico-social/strategie-nationale-de-sante/article/la-strategie-nationale-de-sante-2018-2022

18. Revah-Levy A, Birmaher B, Gasquet I, Falissard B. The Adolescent Depression Rating Scale (ADRS): a validation study. BMC Psychiatry [Internet]. déc 2007 [cité 7 juin 2019];7(1). Disponible sur: https://bmcpsychiatry.biomedcentral.com/articles/10.1186/1471-244X-7-2

19. Hibell B, Guttormsson U, Ahlström S, Balakireva O, Bjarnason T, Kokkevi, et al. The 2007 ESPAD Report: substance use among students in 35 European countries. Stockholm: The Swedish Council for Information on Alcohol and Other Drugs (CAN) : The Pompidou Group at the Council of Europe and the authors; 2009.

20. Beck F, Costes J-M, Legleye S, Spilka S. L'enquête ESCAPAD sur les consommations de drogues des jeunes français: un dispositif original de recueil de l'information sur un sujet sensible. Lavallée P, Rivest L. Québec: Dunod; 2006. (Méthodes d'enquêtes et sondages - Pratiques européennne et nord-américaine).

21. Benarous X, Hassler C, Falissard B, Consoli A, Cohen D. Do girls with depressive symptoms exhibit more physical aggression than boys? A cross sectional study in a national adolescent sample. Child Adolesc Psychiatry Ment Health [Internet]. déc 2015 [cité 1 avr 2020];9(1). Disponible sur: http://www.capmh.com/content/9/1/41

22. Consoli A, Peyre H, Speranza M, Hassler C, Falissard B, Touchette E, et al. Suicidal behaviors in depressed adolescents: role of perceived relationships in the family. Child Adolesc Psychiatry Ment Health. 2013;7(1):8.

23. Thapar A, Collishaw S, Pine DS, Thapar AK. Depression in adolescence. The Lancet. 2012;379(9820):1056-67.

24. McCarty CA, Mason WA, Kosterman R, Hawkins JD, Lengua LJ, McCauley E. Adolescent School Failure Predicts Later Depression Among Girls. J Adolesc Health. août 2008;43(2):180-7.

25. Storksen I, Roysamb E, Holmen TL, Tambs K. Adolescent adjustment and well-being: Effects of parental divorce and distress. Scand J Psychol. févr 2006;47(1):75-84.

26. Harris ES. Adolescent bereavement following the death of a parent: An exploratory study. Child Psychiatry Hum Dev. 1991;21(4):267-81.

27. Duncan GJ, Brooks-Gunn J. Family Poverty, Welfare Reform, and Child Development. Child Dev. janv 2000;71(1):188-96.

28. Najman JM, Hayatbakhsh MR, Clavarino A, Bor W, O'Callaghan MJ, Williams GM. Family Poverty Over the Early Life Course and Recurrent Adolescent and Young Adult Anxiety and Depression: A Longitudinal Study. Am J Public Health. sept 2010;100(9):1719-23.

29. Torikka A, Kaltiala-Heino R, Rimpelä A, Marttunen M, Luukkaala T, Rimpelä M. Self-reported depression is increasing among socio-economically disadvantaged adolescents – repeated cross-sectional surveys from Finland from 2000 to 2011. BMC Public Health. déc 2014;14(1):408.

30. Cairns KE, Yap MBH, Pilkington PD, Jorm AF. Risk and protective factors for depression that adolescents can modify: A systematic review and meta-analysis of longitudinal studies. J Affect Disord. déc 2014;169:61-75.

31. Shrier LA, Harris SK, Sternberg M, Beardslee WR. Associations of Depression, Self-Esteem, and Substance Use with Sexual Risk among Adolescents. Prev Med. sept 2001;33(3):179-89.

32. Soller B, Haynie DL, Kuhlemeier A. Sexual intercourse, romantic relationship inauthenticity, and adolescent mental health. Soc Sci Res. mai 2017;64:237-48.

33. Michel G, Garcia M, Aubron V, Bernadet S, Salla J, Purper-Ouakil D. Adolescent Mental Health and the Choking Game. Pediatrics. févr 2019;143(2):e20173963.

34. Michel G. Psychopathologie des jeux dangereux chez les jeunes : lorsque le plaisir est conditionne par la violence et le risque. 2015;21.

35. Jeux dangereux et pratiques violentes. Ministère de l'éducation Nationale; 2007.

36. Busse H, Harrop T, Gunnell D, Kipping R. Prevalence and associated harm of engagement in self-asphyxial behaviours ('choking game') in young people: a systematic review. Arch Dis Child. déc 2015;100(12):1106-14.

37. Romano H. « JE » DANGEREUX ET PROCESSUS PSYCHIQUES À L'OEUVRE DANS LES PRATIQUES DANGEREUSES. Éditions GREUPP | « Adolescence »; 2011.

38. McDermott B, Baigent M, Chanen A, Fraser L, Graetz B, Newman L, et al. Clinical practice guidelines: depression in adolescents and young adults. Vol. National Health and Medical Research Council (Australia). Melbourne: Beyondblue; 2011.

39. Revah-Levy A, Speranza M, Barry C, Hassler C, Gasquet I, Moro M-R, et al. Association between Body Mass Index and depression: the" fat and jolly" hypothesis for adolescents girls. BMC Public Health. 2011;11(1):649.

40. Kann L, Olsen EO, McManus T, Harris WA, Shanklin SL, Flint KH, et al. Sexual Identity, Sex of Sexual Contacts, and Health-Related Behaviors Among Students in Grades 9-12 - United States and Selected Sites, 2015. Morb Mortal Wkly Rep Surveill Summ Wash DC 2002. 12 2016;65(9):1-202.

41. Marshal MP, Dietz LJ, Friedman MS, Stall R, Smith HA, McGinley J, et al. Suicidality and Depression Disparities Between Sexual Minority and Heterosexual Youth: A Meta-Analytic Review. J Adolesc Health. 1 août 2011;49(2):115-23.

42. Savioja H, Helminen M, Fröjd S, Marttunen M, Kaltiala-Heino R. Sexual experience and self-reported depression across the adolescent years. Health Psychol Behav Med. janv 2015;3(1):337-47.

43. Ueno K. The effects of friendship networks on adolescent depressive symptoms. Soc Sci Res. sept 2005;34(3):484-510.

44. Field T, Diego M, Sanders C. Adolescent depression and risk factors. Adolescence. 2001;36(143):491-8.

45. Barber BL, Eccles JS, Stone MR. Whatever Happened to the Jock, the Brain, and the Princess?: Young Adult Pathways Linked to Adolescent Activity Involvement and Social Identity. J Adolesc Res. sept 2001;16(5):429-55.

46. Fredricks JA, Eccles JS. Is extracurricular participation associated with beneficial outcomes? Concurrent and longitudinal relations. Dev Psychol. juill 2006;42(4):698-713.

47. Mason MJ, Schmidt C, Abraham A, Walker L, Tercyak K. Adolescents' Social Environment and Depression: Social Networks, Extracurricular Activity, and Family Relationship Influences. J Clin Psychol Med Settings. déc 2009;16(4):346-54.

48. Johnson KE, Taliaferro LA. Relationships between physical activity and depressive symptoms among middle and older adolescents: A review of the research literature: Physical Activity and Depressive Symptoms Among Adolescents. J Spec Pediatr Nurs. oct 2011;16(4):235-51.

49. Kandola A, Lewis G, Osborn DP, Stubbs B, Hayes JF. Depressive symptoms and objectively measured physical activity and sedentary behaviour throughout adolescence: a prospective cohort study. Lancet Psychiatry. 2020;7(3):262-71.

50. OMS | Activité physique [Internet]. WHO. World Health Organization; 2020 [cité 11 déc 2020]. Disponible sur: https://www.who.int/dietphysicalactivity/pa/fr/

51. Babiss LA, Gangwisch JE. Sports participation as a protective factor against depression and suicidal ideation in adolescents as mediated by self-esteem and social support. J Dev Behav Pediatr. 2009;30(5):376-84.

52. Panza MJ, Graupensperger S, Agans JP, Doré I, Vella SA, Evans MB. Adolescent Sport Participation and Symptoms of Anxiety and Depression: A Systematic Review and Meta-Analysis. J Sport Exerc Psychol. 21 mai 2020;42(3):201-18.

53. Sanders CE, Field TM, Diego M, Kaplan M. Moderate involvement in sports is related to lower depression levels among adolescents. Adolescence. 2000;35(140):793-7.

54. Maras D, Flament MF, Murray M, Buchholz A, Henderson KA, Obeid N, et al. Screen time is associated with depression and anxiety in Canadian youth. Prev Med. 2015;6.

55. Casiano H, Ma DJK, Katz LY, Rn MJC, Sareen J. Media Use and Health Outcomes in Adolescents: Findings from a Nationally Representative Survey. 2012;6.

56. Liu M, Wu L, Yao S. Dose–response association of screen time-based sedentary behaviour in children and adolescents and depression: a meta-analysis of observational studies. Br J Sports Med. oct 2016;50(20):1252-8.

57. Kaltiala-Heino R, Fröjd S, Marttunen M. Involvement in bullying and depression in a 2-year follow-up in middle adolescence. Eur Child Adolesc Psychiatry. janv 2010;19(1):45-55.

58. Austin PC, Tu JV, Ho JE, Levy D, Lee DS. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. J Clin Epidemiol. avr 2013;66(4):398-407.

59. Walsh CG, Ribeiro JD, Franklin JC. Predicting Risk of Suicide Attempts Over Time Through Machine Learning. Clin Psychol Sci. mai 2017;5(3):457-69.

60. Patton GC, Sawyer SM, Santelli JS, Ross DA, Afifi R, Allen NB, et al. Our future: a Lancet commission on adolescent health and wellbeing. The Lancet. juin 2016;387(10036):2423-78.

61. Leek J, McShane BB, Gelman A, Colquhoun D, Nuijten MB, Goodman SN. Five ways to fix statistics. Nature. nov 2017;551(7682):557-9.

62. Patel CJ, Kerr J, Thomas DC, Mukherjee B, Ritz B, Chatterjee N, et al. Opportunities and Challenges for Environmental Exposure Assessment in Population-Based Studies. Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol. sept 2017;26(9):1370-80.

63. Fan J, Han F, Liu H. Challenges of Big Data analysis. Natl Sci Rev. 1 juin 2014;1(2):293-314.

64. Kotu V, Bala D. Predictive Analytics and Data Mining [Internet]. Elsevier; 2015 [cité 18 déc 2020]. Disponible sur: https://linkinghub.elsevier.com/retrieve/pii/C20140003292

65. Hand DJ. Statistics and data mining: intersecting disciplines. ACM SIGKDD Explor Newsl. 1 juin 1999;1(1):16-9.

66. Smyth P. Data mining: data analysis on a grand scale? Stat Methods Med Res. août 2000;9(4):309-27.

67. Brownlee J. A Tour of Machine Learning Algorithms [Internet]. Machine Learning Mastery. 2019 [cité 18 déc 2020]. Disponible sur: https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

68. Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang J-F, et al. Data mining in healthcare and biomedicine: a survey of the literature. J Med Syst. août 2012;36(4):2431-48.

69. Wiemken TL, Kelley RR. Machine Learning in Epidemiology and Health Outcomes Research. Annu Rev Public Health. 2 avr 2020;41(1):21-36.

70. Hastie T, Tibshirani R, Friedman JH. The Elements ofStatistical LearningData Mining, Inference, and Prediction. Springer Series in Statistics. 2009. 764 p. (Springer Series in Statistics).

71. Noble WS. What is a support vector machine? Nat Biotechnol. déc 2006;24(12):1565-7.

72. Loh W-Y. Classification and Regression Trees. 2011;17.

73. Dimeglio C, Delpierre C, Chauvin P, Lefèvre T. Utilisation des réseaux bayésiens comme technique de fouille de données massives – application à des données de recours aux soins. Rev Francaise Aff Soc. 2017;(4):27-55.

74. Koo CL, Liew MJ, Mohamad MS, Mohamed Salleh AH. A Review for Detecting Gene-Gene Interactions Using Machine Learning Methods in Genetic Epidemiology. BioMed Res Int. 2013;2013:1-13.

75. Molitor J, Papathomas M, Jerrett M, Richardson S. Bayesian profile regression with an application to the National Survey of Children's Health. Biostatistics. 2010;11(3):484-98.

76. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. JAMA. 3 avr 2018;319(13):1317-8.

77. Bellinger C, Mohomed Jabbar MS, Zaïane O, Osornio-Vargas A. A systematic review of data mining and machine learning for air pollution epidemiology. BMC Public Health [Internet]. déc 2017 [cité 15 nov 2018];17(1). Disponible sur: https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-017-4914-3

78. Vallmuur K. Machine learning approaches to analysing textual injury surveillance data: a systematic review. Accid Anal Prev. juin 2015;79:41-9.

79. Alonso SG, de la Torre-Díez I, Hamrioui S, López-Coronado M, Barreno DC, Nozaleda LM, et al. Data Mining Algorithms and Techniques in Mental Health: A Systematic Review. J Med Syst. 21 juill 2018;42(9):161.

80. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine Learning and Data Mining Methods in Diabetes Research. Comput Struct Biotechnol J. 2017;15:104-16.

81. Mooney SJ, Pejaver V. Big Data in Public Health: Terminology, Machine Learning, and Privacy. Annu Rev Public Health. 1 avr 2018;39:95-112.

82. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is Machine Learning? A Primer for the Epidemiologist. Am J Epidemiol. 31 déc 2019;188(12):2222-39.

83. Benke K, Benke G. Artificial Intelligence and Big Data in Public Health. Int J Environ Res Public Health. 10 déc 2018;15(12).

84. Blei DM. Latent Dirichlet Allocation. 2003;30.

85. Gross A, Murthy D. Modeling virtual organizations with Latent Dirichlet Allocation: A case for natural language processing. Neural Netw. oct 2014;58:38-49.

86. Jones T, Doane W. Package 'textmineR'-Functions for Text Mining and Topic Modeling. 2019;

87. Antons D, Kleer R, Salge TO. Mapping the Topic Landscape of JPIM, 1984–2013: In Search of Hidden Structures and Development Trajectories. J Prod Innov Manag. 2016;33(6):726-49.

88. Carson S, Kenny S. Package 'LDAvis'- Interactive Visualization of Topic Models. 2015;

89. Akaike H. A new look at the statistical model identification. IEEE Trans Autom Control. déc 1974;19(6):716-23.

90. Schwarz G. Estimating the Dimension of a Model. Ann Stat. 1978;6(2):461-4.

91. Cole TJ, Bellizzi MC, Flegal KM, Dietz WH. Establishing a standard definition for child overweight and obesity worldwide: international survey. BMJ. 6 mai 2000;320(7244):1240.

92. Bennett DS, Ambrosini PJ, Kudes D, Metz C, Rabinovich H. Gender differences in adolescent depression: Do symptoms differ for boys and girls? J Affect Disord. déc 2005;89(1-3):35-44.

93. McGuinness TM, Dyer JG, Wade EH. Gender differences in adolescent depression. J Psychosoc Nurs Ment Health Serv. déc 2012;50(12):17-20.

94. Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. J Anim Ecol. juill 2008;77(4):802-13.

95. Lampa E, Lind L, Lind PM, Bornefalk-Hermansson A. The identification of complex interactions in epidemiology and toxicology: a simulation study of boosted regression trees. Environ Health. déc 2014;13(1):57.

96. Vivian S. Zhang. Winning data science competitions, presented by Owen Zhang [Internet]. Formation présenté à; 2015 [cité 13 déc 2020]. Disponible sur: https://fr.slideshare.net/ShangxuanZhang/winning-data-science-competitions-presented-by-owen-zhang

97. Hartmann A, Van Der Kooij AJ, Zeeck A. Exploring nonlinear relations: models of clinical decision making by regression with optimal scaling. Psychother Res J Soc Psychother Res. juill 2009;19(4-5):482-92.

98. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. J R Stat Soc Ser B Methodol. janv 1996;58(1):267-88.

99. Engebretsen S, Bohlin J. Statistical predictions with glmnet. Clin Epigenetics. déc 2019;11(1):123.

100. Goldstein BA, Hubbard AE, Cutler A, Barcellos LF. An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. BMC Genet. 14 juin 2010;11:49.

101. Breiman L. Random Forests. Mach Learn. 1 oct 2001;45(1):5-32.

102. Schapire RE. The Boosting Approach to Machine Learning: An Overview. In: Denison DD, Hansen MH, Holmes CC, Mallick B, Yu B, éditeurs. Nonlinear Estimation and Classification [Internet]. New York, NY: Springer New York; 2003 [cité 13 déc 2020]. p. 149-71. (Bickel P, Diggle P, Fienberg S, Krickeberg K, Olkin I, Wermuth N, et al. Lecture Notes in Statistics; vol. 171). Disponible sur: http://link.springer.com/10.1007/978-0-387-21579-2_9

103. Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal. févr 2002;38(4):367-78.

104. Kuhn M. Building Predictive Models in *R* Using the **caret** Package. J Stat Softw [Internet]. 2008 [cité 5 nov 2020];28(5). Disponible sur: http://www.jstatsoft.org/v28/i05/

105. Cohen J. A Coefficient of Agreement for Nominal Scales. Educ Psychol Meas. avr 1960;20(1):37-46.

106. Wasserman L. Discussion: “A significance test for the lasso”. Ann Stat. avr 2014;42(2):501-8.

107. Fisher A, Rudin C, Dominici F. All Models are Wrong, but Many are Useful: Learning a Variable’s Importance by Studying an Entire Class of Prediction Models Simultaneously. ArXiv180101489 Stat [Internet]. 23 déc 2019 [cité 6 nov 2020]; Disponible sur: http://arxiv.org/abs/1801.01489

108. Molnar C. 5.5 Permutation Feature Importance | Interpretable Machine Learning [Internet]. 2018 [cité 13 déc 2020]. Disponible sur: https://christophm.github.io/interpretable-ml-book/feature-importance.html

109. Burzykowski PB and T. 16 Variable-importance Measures | Explanatory Model Analysis [Internet]. 2018 [cité 6 déc 2020]. Disponible sur: https://pbiecek.github.io/ema/featureImportance.html

110. Molnar C, Casalicchio G, Bischl B. iml: An R package for Interpretable Machine Learning. J Open Source Softw. 27 juin 2018;3(26):786.

111. Caye A, Agnew-Blais J, Arseneault L, Gonçalves H, Kieling C, Langley K, et al. A risk calculator to predict adult attention-deficit/hyperactivity disorder: generation and external validation in three birth cohorts and one clinical sample. Epidemiol Psychiatr Sci. 2020;29:e37.

112. Jung JS, Park SJ, Kim EY, Na K-S, Kim YJ, Kim KG. Prediction models for high risk of suicide in Korean adolescents using machine learning techniques. De Luca V, éditeur. PLOS ONE. 6 juin 2019;14(6):e0217639.

113. Miché M, Studerus E, Meyer AH, Gloster AT, Beesdo-Baum K, Wittchen H-U, et al. Prospective prediction of suicide attempts in community adolescents and young adults, using regression methods and machine learning. J Affect Disord. mars 2020;265:570-8.

114. Walsh CG, Ribeiro JD, Franklin JC. Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning. J Child Psychol Psychiatry. déc 2018;59(12):1261-70.

115. Tate AE, McCabe RC, Larsson H, Lundström S, Lichtenstein P, Kuja-Halkola R. Predicting mental health problems in adolescence using machine learning techniques. Mumtaz W, éditeur. PLOS ONE. 6 avr 2020;15(4):e0230389.

116. Foster S, Mohler-Kuo M, Tay L, Hothorn T, Seibold H. Estimating patient-specific treatment advantages in the 'Treatment for Adolescents with Depression Study'. J Psychiatr Res. mai 2019;112:61-70.

117. Pearson R, Pisner D, Meyer B, Shumake J, Beevers CG. A machine learning ensemble to predict treatment outcomes following an Internet intervention for depression. Psychol Med. oct 2019;49(14):2330-41.

118. Olson RS, Cava WL, Mustahsan Z, Varik A, Moore JH. Data-driven advice for applying machine learning to bioinformatics problems. In: Biocomputing 2018 [Internet]. Kohala Coast, Hawaii, USA: WORLD SCIENTIFIC; 2018 [cité 26 nov 2020]. p. 192-203. Disponible sur: https://www.worldscientific.com/doi/abs/10.1142/9789813235533_0018

119. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. juin 2019;110:12-22.

120. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol. déc 2014;14(1):137.

121. Wolpert DH. The Supervised Learning No-Free-Lunch Theorems. In: Roy R, Köppen M, Ovaska S, Furuhashi T, Hoffmann F, éditeurs. Soft Computing and Industry: Recent Applications [Internet]. London: Springer; 2002 [cité 26 nov 2020]. p. 25-42. Disponible sur: https://doi.org/10.1007/978-1-4471-0123-9_3

122. Rahman SA, Walker RC, Lloyd MA, Grace BL, van Boxel GI, Kingma BF, et al. Machine learning to predict early recurrence after oesophageal cancer surgery: Early recurrence after oesophageal cancer surgery. Br J Surg. juill 2020;107(8):1042-52.

123. Zink J, Belcher BR, Imm K, Leventhal AM. The relationship between screen-based sedentary behaviors and symptoms of depression and anxiety in youth: a systematic review of moderating variables. BMC Public Health. déc 2020;20(1):472.

124. Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. BMC Bioinformatics. 11 juill 2008;9(1):307.

125. Hooker G, Mentch L. Please Stop Permuting Features: An Explanation and Alternatives. ArXiv190503151 Cs Stat [Internet]. 1 mai 2019 [cité 26 nov 2020]; Disponible sur: http://arxiv.org/abs/1905.03151

126. Molnar C, Casalicchio G, Bischl B. Interpretable Machine Learning -- A Brief History, State-of-the-Art and Challenges. ArXiv201009337 Cs Stat [Internet]. 19 oct 2020 [cité 26 nov 2020]; Disponible sur: http://arxiv.org/abs/2010.09337

127. Salami D, Sousa CA, Martins M do RO, Capinha C. Predicting dengue importation into Europe, using machine learning and model-agnostic methods. Sci Rep. déc 2020;10(1):9689.

128. Friedman JH, Popescu BE. Predictive Learning via Rule Ensembles. Ann Appl Stat. 2008;2(3):916-54.

129. Greenwell BM, Boehmke BC, McCarthy AJ. A Simple and Effective Model-Based Variable Importance Measure. ArXiv180504755 Cs Stat [Internet]. 12 mai 2018 [cité 26 nov 2020]; Disponible sur: http://arxiv.org/abs/1805.04755

130. Oh S. Feature Interaction in Terms of Prediction Performance. Appl Sci. 29 nov 2019;9(23):5191.

131. Molnar C. 5.6 Global Surrogate | Interpretable Machine Learning [Internet]. 2018 [cité 14 déc 2020]. Disponible sur: https://christophm.github.io/interpretable-ml-book/global.html

132. Hastie DI, Liverani S, Azizi L, Richardson S, Stücker I. A semi-parametric approach to estimate risk functions associated with multi-dimensional exposure profiles: application to smoking and lung cancer. BMC Med Res Methodol. déc 2013;13(1):129.

133. Mattei F, Liverani S, Guida F, Matrat M, Cenée S, Azizi L, et al. Multidimensional analysis of the effect of occupational exposure to organic solvents on lung cancer risk: the ICARE study. Occup Env Med. 2016;oemed-2015.

134. Sarra A, Fontanella L, Di Zio S. Identifying Students at Risk of Academic Failure Within the Educational Data Mining Framework. Soc Indic Res. nov 2019;146(1-2):41-60.

135. Liverani S, Hastie DI, Azizi L, Papathomas M, Richardson S. PReMiuM: An R Package for Profile Regression Mixture Models using Dirichlet Processes. ArXiv13032836 Stat [Internet]. 12 mars 2013 [cité 10 sept 2019]; Disponible sur: http://arxiv.org/abs/1303.2836

136. Papathomas M, Molitor J, Hoggart C, Hastie D, Richardson S. Exploring Data From Genetic Association Studies Using Bayesian Variable Selection and the Dirichlet Process: Application to Searching for Gene × Gene Patterns: Searching for Gene × Gene Patterns. Genet Epidemiol. sept 2012;36(6):663-74.

137. Coker E, Gunier R, Bradman A, Harley K, Kogut K, Molitor J, et al. Association between Pesticide Profiles Used on Agricultural Fields near Maternal Residences during Pregnancy and IQ at Age 7 Years. Int J Environ Res Public Health. 9 mai 2017;14(5):506.

138. Armoni Domany K, Hossain MM, Nava-Guerra L, Khoo MC, McConnell K, Carroll JL, et al. Cardioventilatory Control in Preterm-born Children and the Risk of Obstructive Sleep Apnea. Am J Respir Crit Care Med. 15 juin 2018;197(12):1596-603.

139. Ismaili OA, Lemaire V, Cornuéjols A. Classification à base de clustering : ou comment décrire et prédire simultanément. :6.

140. Spilka S, Le Nézet O. Alcool, tabac et cannabis durant les« années lycée ». OFDT; 2013.

141. Spilka S, Ehlinger V, Le Nézet O, Pacoricona D, Ngantcha M, Godeau E. Alcool, tabac et cannabis en 2014, durant les « années collège ». OFDT; 2015.

142. Dubet F, Martucceli D. À l'école. Sociologie de l'expérience scolaire. Paris Seuil. 1996.

143. Nicole C. Psychopathologie de la scolarité. Elsevier Masson. 2012. 432 p.

144. Barcella W, Iorio MD, Baio G, Malone-Lee J. Variable selection in covariate dependent random partition models: an application to urinary tract infection: Variable selection in covariate dependent random partition models: an application to urinary tract infection. Stat Med. 15 avr 2016;35(8):1373-89.

145. Barcella W, De Iorio M, Baio G. A comparative review of variable selection techniques for covariate dependent Dirichlet process mixture models. Can J Stat. sept 2017;45(3):254-73.

146. Spilka S, Janssen E. DETECTION DES USAGES PROBLEMATIQUES DE CANNABIS : LE CANNABIS ABUSE SCREENING TEST (CAST). :9.

147. Lemmens JS, Valkenburg PM, Peter J. Development and Validation of a Game Addiction Scale for Adolescents. Media Psychol. 26 févr 2009;12(1):77-95.

148. Enquête TNS-Sofres, « Jeux dangereux en milieux scolaire et extra-scolaire : le point de vue des parents, la pratique des enfants » [Internet]. TNS-Sofres; 2007. Disponible sur: www.tns-healthcare.fr/fichiers/etudes/00000066.pdf

149. Brausch AM, Decker KM, Hadley AG. Risk of Suicidal Ideation in Adolescents with both SelfAsphyxial RiskTaking Behavior and NonSuicidal SelfInjury. 2011;11.

150. Centers for Disease Control and Prevention (CDC). « Choking game » awareness and participation among 8th graders--Oregon, 2008. MMWR Morb Mortal Wkly Rep. 15 janv 2010;59(1):1-5.

151. Dake JA, Price JH, Kolm-Valdivia N, Wielinski M. Association of Adolescent Choking Game Activity With Selected Risk Behaviors. Acad Pediatr. nov 2010;10(6):410-6.

152. Ramowski SK, Nystrom RJ, Rosenberg KD, Gilchrist J, Chaumeton NR. Health Risks of Oregon Eighth-Grade Participants in the "Choking Game": Results From a Population-Based Survey. Pediatrics. 1 mai 2012;129(5):846-51.

153. Bernadet S, Purper-Ouakil D, Michel G. Typologie des jeux dangereux chez des collégiens : vers une étude des profils psychologiques. Ann Méd-Psychol Rev Psychiatr. nov 2012;170(9):654-8.

154. Allahverdipour H, Bazargan M, Farhadinasab A, Moeini B. Correlates of video games playing among adolescents in an Islamic country. 2010;7.

155. Boers E, Afzali MH, Newton N, Conrod P. Association of Screen Time and Depression in Adolescence. JAMA Pediatr. 1 **sept** 2019;173(9):853.

156. Primack BA, Swanier B, Georgiopoulos AM, Land SR, Fine MJ. Association Between Media Use in Adolescence and Depression in Young Adulthood: A Longitudinal Study. Arch Gen Psychiatry. 1 févr 2009;66(2):181.

157. Eime RM, Young JA, Harvey JT, Charity MJ, Payne WR. A systematic review of the psychological and social benefits of participation in sport for children and adolescents: informing development of a conceptual model of health through sport. Int J Behav Nutr Phys Act. 2013;10(1):98.

université PARIS-SACLAY | ÉCOLE DOCTORALE Santé Publique (EDSP)

Titre : Méthodes de fouilles de données en épidémiologie psychiatrique : application à l'analyse des facteurs et marqueurs de risque de la symptomatologie dépressive à l'adolescence.

Mots clés : Fouilles de données ; apprentissage automatique ; dépression ; adolescence

Résumé : L'adolescence est une période de vulnérabilité pour la dépression, sur le plan psychologique et biologique. La littérature sur la dépression à l'adolescence est très fournie sur ses facteurs de risque et de protection ainsi que sur les différentes manifestations externalisées pouvant servir de signe d'appel. Cependant, les modèles de prédiction du risque restent peu performants. La recherche systématique et approfondie des combinaisons entre marqueurs/facteurs de risque pourrait être un moyen d'améliorer ces modèles. Les techniques issues des méthodes de « fouille de données » (data mining, machine Learning DMML) semblent de plus en plus utilisées sur des problématiques similaires. Ce travail de thèse va s'intéresser à l'application des méthodes issues du DMML à la dépression durant l'adolescence. Dans ce contexte, l'objectif sera i) de cartographier l'utilisation réelle de ces méthodes en épidémiologie et santé publique ii) d'analyser les patterns d'interactions entre les facteurs/marqueurs de risque de la dépression à l'adolescence afin de développer de nouvelles pistes utiles dans le repérage de cette population.

En premier lieu, une analyse bibliométrique de Medline, sera réalisée afin de quantifier l'essor des méthodes issues du DMML en santé publique et épidémiologie et d'en caractériser les domaines d'application majeurs. Dans un second temps, une comparaison de l'apport de deux méthodes de classification quant à leur capacité à modéliser le risque de dépression : ensemble d'arbres par régression boostée, des forêts aléatoires par rapport à une régression logistique LASSO sans interaction sera réalisée. Pour finir, une méthode de partitionnement supervisée, appelée « Régression sur profil», sera utilisée pour créer des clusters d'adolescents à partir des variables explicatives de la dépression et de la dépression. Les données issues de l'enquête transversale en milieu scolaire «Processus d'adolescence» seront utilisées. Elle inclut, 15235 adolescents, répondant à un auto-questionnaire anonyme contenant la dépression via l'Adolescent Depression Rating Scale et les variables explicatives de la dépression présentes dans l'enquête. Cette thèse a montré les intérêts et les difficultés quant à l'utilisation des méthodes issues du DMML pour la recherche d'associations pertinentes en épidémiologie psychiatrique.

Title : Data mining methods in psychiatric epidemiology: application on the analysis of risk factors in depressive symptoms at adolescence

Keywords : Data mining ; Machine Learning ; depression ; adolescence

Abstract: Adolescence is a vulnerable period for depression, both psychologically and biologically. The literature on depression in adolescence is very extensive on risk and protective factors and on the various externalized manifestations that can serve as warning sign. However, prediction models remain poorly performing. Systematic and in-depth research into the combinations of risk factors/markers could improving these models. Techniques derived from data mining/Machine Learning methods (DMML) now seem to be more and more used on similar issues. This work will focus on the application of DMML methods to depression during adolescence. In this context, the objective will be i) to map the actual use of these methods in epidemiology and public health ii) to analyze the associations between risk factors/markers of depression in adolescence in order to develop new useful leads in the identification of this population. First, a bibliometric analysis of Medline will be conducted in order to quantify the development of DMML methods in public health and epidemiology and to characterize their major fields of application.

Secondly, a comparison of the contribution of two classification methods in terms of their capacity to model the risk of depression: boosted regression trees, random forests compared to a logistic LASSO regression without interaction will be carried out. Finally, a supervised partitioning method, called «Bayesian Profile regression", will be used to create clusters of adolescents from the explanatory variables of depression and depression. Data from the cross-sectional school survey "Processus d'adolescence" will be used. It includes 15235 adolescents, responding to an anonymous self-administered questionnaire containing depression via the Adolescent Depression Rating Scale and the explanatory variables for depression present in the survey. This work showed the interests and difficulties of DMML to analysis relevant associations in psychiatric epidemiology.