UNIVERSITÉ DE NANTES

UNIVERSITE DE NANTES

_____

FACULTE DE MEDECINE

_____

Année : 2020 N° 2020-07

THESE

pour le

DIPLOME D'ETAT DE DOCTEUR EN MEDECINE

DIPLOME D'ETUDES SPECIALISEES DE

SANTE PUBLIQUE ET MEDECINE SOCIALE

par

Delphine MOURET

née le 05 mai 1990 à Lavaur

_____

Présentée et soutenue publiquement le 28 janvier 2020

_____

Etude de l'association entre l'exposition aux polluants organiques persistants (POPs) et l'endométriose, à l'aide d'algorithmes d'apprentissage statistique.

Exploitation des données du projet ENDOTOX : étude cas-témoins réalisée en Pays de la Loire entre 2013 et 2015

_____

Président : Monsieur le Professeur Pierre-Antoine GOURRAUD
Directeur de thèse : Monsieur le Docteur German CANO-SANCHO

UNIVERSITÉ DE NANTES

I. REMERCIEMENTS

A Monsieur le Professeur Pierre-Antoine Gourraud, pour s'être intéressé à cette étude, pour m'avoir fait confiance et m'avoir fait l'honneur d'accepter de présider ce jury.

A Monsieur le Professeur Stéphane Ploteau, pour ses différents travaux sur l'endométriose sur lesquels je me suis grandement appuyée et sans lesquels je n'aurais pu effectuer ce travail, et pour avoir accepté d'être membre de ce jury.

A Monsieur le Docteur Jean-Pascal Fournier, pour avoir présenté un intérêt à ce sujet et avoir accepté d'être membre de ce jury

A Monsieur le Docteur German Cano-Sancho, pour m'avoir accueilli en stage, pour ta gentillesse, ta patience, ton aide précieuse et d'avoir accepté d'être mon directeur de thèse.

A Monsieur le Docteur Jean-Philippe Antignac pour avoir supervisé ce travail, pour avoir pris le temps de me relire et de me donner de bons conseils pour ce travail.

A l'équipe du LABoratoire d'Etudes des Résidus et Contaminants dans les Aliments (LABERCA), UMR 1329 Inra-Oniris, pour votre accueil et votre bonne humeur.

Au futur Docteur Komodo Matta, pour m'avoir accompagné durant mon stage au LABERCA et pour avoir valorisé ce travail par un article soumis.

A Madame le Professeur Evelyne Vigneau, et à Madame le Docteur Véronique Cariou, pour m'avoir aidé sur le plan méthodologique, et m'avoir apporté votre soutien.

A Madame le Professeur Leila Moret, pour m'avoir soutenu tout au long de mon internat de santé publique.

A Monsieur le Docteur Brice Leclère, pour sa pédagogie.

A toutes les équipes qui m'ont reçu en stage tout au long de mes études.

A l'ensemble des personnes qui m'ont accompagné, à mes proches et ma famille qui m'ont supporté tout au long de cette première décennie d'études.

UNIVERSITÉ DE NANTES

II. TABLE DES MATIÈRES

I. REMERCIEMENTS 2

II. TABLE DES MATIÈRES 3

III. INTRODUCTION 6

IV. CONTEXTE 7

1. Environnement et santé 7

a) Notion d'« exposome » 7

b) Politiques publiques nationales 7

c) Epidémiologie et facteurs environnementaux 8

2. Endométriose 8

a) Définition 8

b) Facteurs de risques environnementaux 9

c) Origine des polluants organiques persistants associés à l'endométriose 10

3. Projet ENDOTOX : Endométriose et environnement en France 12

4. Limites des approches statistiques en épidémiologie environnementale 13

5. Intelligence artificielle, dont l'apprentissage automatique, en épidémiologie et en santé publique 14

a) Émergence de méthodes statistiques basées sur l'intelligence artificielle 14

b) Définition de l'apprentissage automatique 14

c) Applications de ces méthodes dans le domaine de la santé 15

6. Revue de la littérature des méthodes statistiques d'intérêt pour aborder l'effet « cocktail » en épidémiologie environnementale 16

7. Objectifs 17

V. MATERIEL 18

1. Présentation de la base de données : projet ENDOTOX 18

2. Logiciels 19

VI. MÉTHODES 20

1. Sélection des modèles 20

2. Schéma de modélisation et d'application des algorithmes 21

a) Préparation des données 22

b) Séparation du jeu de données en un jeu d'entraînement et jeu de test 22

c) Entraînement des modèles et calibration des paramètres 22

d) Ajustement et évaluation des différents modèles sur le jeu de test 23

3. Description et calibration des modèles sélectionnés 24

a) Réseaux de neurones artificiels 24

i. Description du modèle 24

ii. Calibration du modèle 26

iii. Estimation de l'importance des variables 27

b) Support Vector Machine 27

i. Description 27

ii. Calibration du modèle 29

iii. Estimation de l'importance des variables 30

c) Boosting trees 30

i. Description 30

ii. Calibration du modèle 31

iii. Estimation de l'importance des variables 32

VII. RESULTATS 32

1. Réseaux de neurones artificiels 32

a) Optimisation et choix des paramètres 32

b) Importance des variables 34

2. Support Vector Machine 35

a) Optimisation et choix des paramètres 35

b) Importance des variables 35

3. « Boosting Trees » 37

a) Optimisation et choix du paramètres 37

b) Importance des variables 37

4. Comparaison des trois modèles 39

VIII. DISCUSSION 41

1. Polluants organiques persistants associés à l'endométriose et performance des modèles utilisés 41

2. Forces de l'étude 42
a) Application des modèles dans ce cas d'étude 42
b) Applications de ces modèles à d'autres études 43

3. Limites de l'étude 43
a) Liées aux données 43
b) Liées aux modèles 44
c) Liées à l'interprétabilité 45

4. Perspectives 45
a) Epidémiologie environnementale 45
b) Ouverture à la santé au sens général 46

IX. CONCLUSION 49

X. BIBLIOGRAPHIE 50

XI. ANNEXES 57

1. Tableau A1 : Revue de la littérature 57

2. Tableau A2 : Régression logistique avec OR et p-value de chaque polluant 60

3. Tableau A3 : Coefficients obtenus avec la régression logistique pénalisée Elastic-Net (Cano-Sancho, 2017) 62

4. Tableau A4 : Polluants de la base de données ENDOTOX 63

5. Tableau A5 : Liste des packages et fonctions utilisées dans le logiciel R 65

6. Interprétation d'un arbre de classification simple avec la base ENDOTOX 66

7. Tableau A6 : Résultats coefficients (poids) des ANN 68

XII. RÉSUMÉ 69

1. Résumé en français (500 mots) 69

2. Summary in english (500 words) 70

XIII. SIGNATURES 72

III. INTRODUCTION

Les facteurs environnementaux contribuent à l'étiologie d'un grand nombre de pathologies. Les études de biosurveillance montrent que nous sommes exposés à une multitude de produits chimiques, mais la caractérisation de l'impact sur la santé de ces mélanges complexes de polluants reste un défi majeur.

L'endométriose est une pathologie gynécologique, inflammatoire et hormono-dépendante, considérée comme une pathologie fréquente, car elle touche entre 5 à 10% des femmes en âge de procréer. Elle est caractérisée par une migration anormale du tissu utérin (endomètre) à l'extérieur de la cavité utérine. L'étiologie de cette pathologie reste encore imparfaitement caractérisée mais apparait dans tous les cas multi-causale, avec une part hormonale ou œstrogèno-dépendante, génétique et environnementale.

Plusieurs études ont montré des relations entre des polluants organiques persistants comme les dioxines, les polychlorobiphényles (PCBs) ou les pesticides organochlorés (OCPs) et l'endométriose. Cependant l'évidence épidémiologique reste globalement peu concluante, et les méthodes statistiques utilisées présentent de multiples limites. Une des limites des approches statistiques classiques usuellement mises en œuvre en épidémiologie environnementale, de type régression linéaire et régression logistique, est de ne pas tenir compte de l'effet « cocktail » de polluants et de leurs interactions, ce qui devrait être au contraire considéré dans des scénarios réalistes. La stabilité des modèles de régression linéaires se trouve compromise, notamment en raison de la forte multicollinéarité et de la redondance des variables chimiques en épidémiologie environnementale. Il existe aujourd'hui un nombre croissant de méthodes alternatives dans le domaine de l'apprentissage automatique (« machine learning ») pour pouvoir pallier à ces contraintes. Ces méthodes sont notamment utilisées pour la prédiction de pathologies et de leurs évolutions, ou pour la recommandation de traitements personnalisés, l'aide au diagnostic, …. Une grande partie de la littérature en santé sur l'intelligence artificielle analyse des données d'imagerie médicale, d'électrodiagnostic et de génétique. Les principaux types de pathologies présents dans la littérature, étudiée avec ces méthodes sont notamment : les cancers, les pathologies neurologiques, cardio-vasculaire, urologique, etc.

Ces méthodes constituent donc une alternative prometteuse aux modèles conventionnels pour étudier ces données aux structures complexes. Elles sont toutefois relativement récentes, et leur potentiel semble encore partiellement inexploité et que des développements de ces approches permettant une extraction facilitée de connaissance peuvent être attendus dans les prochaines années.

L'objectif de ce travail ici, est d'étudier l'association entre les niveaux d'exposition internes de polluants et la présence d'endométriose à l'aide de trois approches « multi-polluants » d'apprentissage automatique : les réseaux de neurones, les machines à vecteurs de support, et les « Boosting Trees »

IV. CONTEXTE

1. Environnement et santé

a) Notion d'« exposome »

De multiples études ont maintenant prouvé que l'Homme est exposé à des mélanges complexes de substances chimiques durant toute sa vie. Ce constat s'inscrit aujourd'hui dans le concept d' « exposome » introduit en 2005 par l'épidémiologiste Christopher Wild (Wild 2012). Le concept d'exposome vise à prendre en compte la totalité des expositions environnementales, subie par un être humain depuis sa conception, en le croisant avec les caractéristiques de son génome pour mieux comprendre l'étiologie des pathologies humaines (1). L'exposition à certains contaminants chimiques environnementaux, à certaines concentrations et durant des fenêtres de sensibilité particulières (période périnatale, puberté...), peut avoir un impact sur certains processus biologiques internes et à terme contribuer à l'apparition et/ou le développement de certaines maladies. Il apparaît donc nécessaire d'évaluer l'association entre ces expositions et la modification de paramètres de santé, ce qui est l'objectif principal de l'épidémiologie environnementale.

b) Politiques publiques nationales

Au cours du 20ème siècle, les conséquences sanitaires de l'amiante, et le temps nécessaire pour les gérer en France et dans le monde, ont marqué un tournant dans la prise en compte par les pouvoirs publics de la santé environnementale. À cet égard, 2004 constitue une année charnière, avec l'adoption de la Charte de l'environnement et du premier plan national santé-environnement (PNSE), qui témoigne de la volonté du gouvernement de réduire autant que possible et de façon la plus efficace les impacts des facteurs environnementaux sur la santé afin de permettre à chacun de vivre dans un environnement favorable à la santé. Sous l'égide des ministères en charge de l'Environnement et de la Santé, la quatrième édition de ce plan est actuellement élaborée en partenariat avec les partenaires intéressés à la santé environnementale (2). Afin de prendre en compte les enjeux territoriaux, le PNSE fait également l'objet de déclinaisons locales à travers les plans régionaux santé-environnement conduits par les Agences Régionales de Santé (ARS). En parallèle, d'autres plans thématiques (Stratégie nationale sur les perturbateurs endocriniens, plan Écophyto, plan Micropolluants, plan national Chlordécone, plan Écoantibio, etc.) ont vu le jour et sont régulièrement actualisés, favorisant le dialogue avec les partenaires lors de leur conception (3).

c) Epidémiologie et facteurs environnementaux

L'épidémiologie étudie la répartition et les déterminants des maladies dans la population. Elle procède par des enquêtes, et permet d'estimer le risque de devenir malade sur une période donnée, et l'augmentation (ou la diminution) du risque associé à nos gènes, nos comportements, ou notre environnement. L'épidémiologie a connu un développement rapide lors de la deuxième moitié du XXe siècle, avec l'identification des principaux facteurs de risque des maladies cardio-vasculaires et de nombreux cancers. Les principaux facteurs de risque comportementaux des maladies chroniques ont été découverts, et la discipline se heurte à l'exploration plus complexe des déterminants génétiques ou environnementaux (4).
De plus en plus d'études et de cohortes au niveau international et national ont vu le jour récemment, étudiant notamment l'impact des facteurs environnementaux comme les pesticides, perturbateurs endocriniens, …. Pour n'en citer que certaines en France : la cohorte Elfe de Santé publique France (5), ayant pour but de mieux connaître les facteurs (environnement, entourage familial, conditions de vie…) qui peuvent avoir une influence sur le développement physique et psychologique de l'enfant, sa santé et sa socialisation, l'étude Esteban de Santé Publique France également (6), visant notamment à mesurer notre exposition à certaines substances de l'environnement, à mieux connaître notre alimentation et notre activité physique et à mesurer l'importance de certaines maladies chroniques dans la population. La cohorte AGRIculture et CANcer (AGRICAN) (7), visant à savoir quels sont les éventuels secteurs professionnels, tâches et nuisances (dont les expositions directes et indirectes aux pesticides) à risque de cancers chez la population agricole.

2. Endométriose

a) Définition

L'endométriose est une pathologie gynécologique, inflammatoire et hormono-dépendante, considérée comme une pathologie fréquente, car elle touche entre 5 à 10% des femmes en âge de procréer. Elle est caractérisée par une migration anormale du tissu utérin (endomètre) à l'extérieur de la cavité utérine. Cette pathologie cause de nombreux symptômes, souvent non-spécifiques, comme par exemple, des douleurs pelviennes chroniques intenses, des troubles des menstruations (dysménorrhées), des troubles de la fertilité, des douleurs pendant les rapports sexuels (dyspareunie). Des lésions peuvent aussi être retrouvées chez des femmes totalement asymptomatiques. Au niveau histologique, il existe trois types d'endométriose : l'endométriose pelvienne, ovarienne et profonde et peuvent se manifester en même temps chez une femme (8) (9).
La physiopathologie de cette maladie reste hypothétique, et plusieurs théories existent. La théorie la plus ancienne et la plus acceptée est celle de la menstruation rétrograde : le reflux rétrograde de

cellules endométriales dans la cavité péritonéale par les trompes lors des règles (10). Les cellules endométriales adhèrent ensuite au péritoine et se développent. Cependant, cette théorie n'explique pas certaines localisations à distance (cerveau, œil, …), ni certaines formes de la maladie, présente chez des femmes ayant une aplasie de l'utérus et du vagin. Par ailleurs, alors que les cliniciens estiment que 90% des femmes présentent des saignements rétrogrades, seules 10% développent des lésions d'endométriose (11).

L'étiologie de cette pathologie reste encore imparfaitement caractérisée mais apparait dans tous les cas multi-causale, avec une part hormonale ou œstrogèno-dépendante, génétique et environnementale, mais aussi des dysfonctions du système immunitaire et des réactions pro-inflammatoires.

L'endométriose est une pathologie oestrogénodépendante et réagit à la suppression de la production hormonale ovarienne (12). L'évolution de la maladie semble en rapport avec une expression aberrante de diverses molécules intervenant dans la croissance et la différenciation cellulaire, comprenant les métalloprotéases, l'aromatase, les facteurs de croissance, certaines oncoproteines (c-myc et c-erbB-2) et les récepteurs stéroïdiens (8). Les résultats de plusieurs études montrent également la contribution de la génétique dans la physiopathologie de l'endométriose (13) (14). Un antécédent d'endométriose sévère chez une parente du premier degré pourrait ainsi multiplier par 6 le risque de développer la pathologie (8).

b) Facteurs de risques environnementaux

Il est maintenant prouvé qu'il existe une relation entre certains polluants, notamment de type perturbateurs endocriniens, et les pathologies gynécologiques (15). Certaines études ont montré des relations entre des polluants organiques persistants comme les dioxines, les polychlorobiphényles (PCBs), les retardateurs de flamme bromés ou les pesticides organochlorés (OCPs) et l'endométriose. Cependant l'évidence épidémiologique reste globalement peu concluante (16). Certaines études sur les primates et les rongeurs suggèrent que l'exposition à certaines dioxines est associée à l'endométriose, mais de telles associations sont incohérentes avec les études sur l'homme et avec d'autres expositions aux organochlorés (17) (16). L'incohérence des études épidémiologiques est probablement due à des problèmes méthodologiques et à l'hétérogénéité des populations (18) (16) mais aussi aux différences de quantification des concentrations chimiques en fonction des laboratoires (19) (20).

Une récente méta-analyse de 2019 (20) a étudié les associations entre les composés chimiques organochlorés, et les estimations étaient statistiquement significatives avec un OR (IC 95%) de 1,65 (1,14; 2,39) pour les dioxines (n = 10 études), 1,70 (1,20; 2,39) pour les PCB (n = 9 études) et 1,23 (1,13; 1,36) pour les OCP (n = 5 études). Cependant, ces résultats doivent être considérés avec prudence, étant donné l'hétérogénéité considérable et la petite taille de l'effet estimé. De plus, il n'existe à ce jour aucun modèle statistique satisfaisant pour étudier l'effet de ces mélanges de

polluants. Cette méta-analyse recommande donc de poursuivre les études épidémiologiques sur ce sujet, avec une méthodologie rigoureuse et l'utilisation d'approches alternatives (modélisation, machine learning), capable de gérer les expositions environnementales à des mélanges chimiques. Etant donné l'impact économique et sociétal de l'endométriose, il est important de continuer les recherches afin de pouvoir aider les décideurs à mettre en place des mesures préventives.

c) Origine des polluants organiques persistants associés à l'endométriose

Dioxines :

Le mot « dioxine » est un terme générique regroupant deux hydrocarbures aromatiques polycycliques chlorés (HAPC) : les dioxines (polychlorodibenzodioxines ou PCDD) et furanes (polychlorodibenzofuranes ou PCDF). Les dioxines sont produites involontairement au cours de la plupart des processus de combustions naturelles et industrielles et en particulier de procédés faisant intervenir de fortes températures (incinération, métallurgie…). Les principales sources d'exposition sont les incinérateurs d'ordures ménagères d'ancienne génération. On les retrouve aussi dans un grand nombre de procédés de fabrication : la métallurgie du cuivre et de l'acier, le blanchiment au chlore des pâtes à papier, la production de certains herbicides et pesticides. Les dioxines sont également émises par les voitures ainsi que lors de la combustion du charbon et du bois. Elles peuvent aussi apparaître au cours de phénomènes naturels, comme les éruptions volcaniques ou les feux de forêts.
Les dioxines sont des composés chimiques peu volatils, peu solubles dans l'eau mais avec une stabilité chimique et métabolique importante, ce qui explique leur forte persistance dans l'environnement. Les dioxines présentent également une affinité pour les graisses (ou lipophilie) qui leur permet de s'accumuler dans les tissus gras de l'organisme pendant une période d'environ sept ans. Chez l'homme, l'exposition moyenne des populations se fait majoritairement par voie alimentaire, en particulier par l'ingestion des graisses animales contenues dans les produits laitiers, viandes, poissons (21) (22) (23).

PolyChloroBiphényls (PCBs) :

Les PCB sont des produits organiques de synthèse aromatiques et chlorés. Ces composés ont été produits en masse à partir des années 1930 pour être utilisés principalement comme isolants dans les équipements électriques mais aussi comme solvants, plastifiants et additifs de fluides hydrauliques ou caloporteurs. Ces composés ont vu leur utilisation progressivement restreinte puis stoppée en raison de leur nocivité (en France, interdiction de production et d'utilisation depuis 1987). Néanmoins, ces polluants organiques persistants sont encore très présents dans l'environnement (tous milieux confondus avec une présence notable dans les sédiments) et peuvent être transportés sur de grandes

distances dans l'environnement (présence en arctique par exemple). Les émissions actuelles de PCB sont diffuses et d'origine anthropique. Ces sources résultent essentiellement de l'usage des derniers appareils contenant encore des PCB (vieux transformateurs susceptibles de fuir, d'exploser ou de bruler), des peintures anciennes, des sites de traitements de matériels contenant des PCB, des zones polluées aux PCB, etc. (24).

Les retardateurs de flamme bromés :

Les retardateurs de flamme bromés (RFB) sont des mélanges de produits chimiques produits par l'homme et ajoutés à des produits variés, notamment pour une utilisation industrielle, afin de les rendre moins facilement inflammables. Ils sont couramment utilisés dans des plastiques, des textiles et des équipements électriques et électroniques.
Il existe cinq catégories principales de RFB, dont la liste figure ci-dessous, avec leurs usages courants :

- Les diphényléthers polybromés (PBDE) – plastiques, textiles, moulages électroniques, circuits
- L'hexabromocyclododécane (HBCDD) – isolation thermique dans l'industrie du bâtiment
- Le tétrabromobisphénol A (TBBPA) et autres phénols – cartes de circuits imprimés, thermoplastiques (principalement dans les téléviseurs)
- Les biphényles polybromés (PBB) – appareils ménagers, textiles, mousses plastiques

Ces types de retardateurs ont été commercialisés sous forme de mélanges techniques par différentes marques commerciales. Dans l'Union européenne (UE), l'utilisation de certains RFB est interdite ou limitée. Cependant, en raison de leur persistance dans l'environnement, il subsiste des inquiétudes concernant les risques que ces produits chimiques présentent pour la santé publique. Les produits traités aux RFB, actuellement en cours d'utilisation ou sous forme de déchets, relâchent des substances qui s'infiltrent dans l'environnement et contaminent l'air, le sol et l'eau. Ces contaminants peuvent ensuite s'introduire dans la chaîne alimentaire, où ils sont présents principalement dans des aliments d'origine animale, tel que le poisson, la viande, le lait et les produits dérivés (25).

Pesticides organochlorés :

Les pesticides organochlorés ont été massivement utilisés partout dans le monde, comme insecticides de contact et dans une moindre mesure comme fongicides et acaricides. Leur spectre d'action est donc très large. En raison d'une forte utilisation ubiquitaire et répétée, leur efficacité s'est amenuisée progressivement, obligeant les utilisateurs à augmenter les doses appliquées. Comme ces composés sont très persistants dans l'environnement et s'accumulent facilement dans les graisses, ils font l'objet d'une bioaccumulation le long de la chaîne alimentaire. L'homme étant à la fin de la chaîne alimentaire consommera donc le plus de pesticides organochlorés. Aujourd'hui leur usage est interdit dans de nombreux pays par le biais d'accords internationaux, tels que la convention de Stockholm. (26) (27).

3. Projet ENDOTOX : Endométriose et environnement en France

Le Projet ENDOTOX est un des projets récemment conduit au sein du LABERCA en collaboration avec le service de chirurgie gynécologique du Centre Hospitalier Universitaire de Nantes dans le cadre de la thèse du Pr Stéphane Ploteau, qui a donné lieu à une description détaillée d'un ensemble de 80 polluants persistants dans le sérum et le tissu adipeux d'un échantillon de population présentant ou non une pathologie de type endométriose. Les données ont été recueillies dans le cadre d'une étude cas-témoins réalisée dans la Région Pays-de-Loire entre 2013 et 2015. Au total, 99 femmes ont été incluses dans l'étude, les cas (n = 55) étaient des femmes adultes, âgées de 18 à 45 ans, avec un diagnostic chirurgical d'endométriose profonde avec/sans endométriose ovarienne. Les sujets témoins (n = 45) étaient des femmes adultes d'âge similaire, consultant pour un autre problème gynécologique mais sans endométriose détectée à la chirurgie et sans aucun symptôme clinique de la maladie comme la douleur pelvienne chronique, la dysménorrhée, la dyspareunie ou l'histoire de l'infertilité. Des informations détaillées ont été recueillies sur le diagnostic, les variables anthropométriques et d'autres facteurs potentiellement associés à l'exposition aux polluants organique persistants, y compris l'âge, l'indice de masse corporelle (IMC), la durée de l'allaitement et la parité. Pour les cas et les témoins, des échantillons (2 ml) de tissu adipeux pariétal (graisse sous cutanée) et de tissu adipeux profond (graisse omentale) ont été prélevés pendant la chirurgie. Un échantillon de sérum (20 ml) a également été prélevé le jour avant la chirurgie. Le tissu adipeux est une représentation au long terme d'une exposition aux contaminants comparée au sérum, ce qui est plus pertinent car le temps entre l'exposition et le développement de l'endométriose est estimé à environ 7 à 10 ans (8) (28) . Une analyse préliminaire des données a été effectuée en utilisant la régression logistique binomiale et multinomiale pour déterminer les associations entre l'endométriose profonde, l'endométriose ovarienne et les polluants considérés individuellement, mettant en évidence une concentration de polluant (dans le tissu adipeux) plus élevée chez les femmes atteintes d'endométriose. Les résultats montrent une association significative entre la présence d'endométriose infiltrante profonde et plusieurs polluants du tissu adipeux : le 1.2.3.7.8 pentachlorodibenzodioxine (PeCDD), l'octachlorodibenzofurane (OCDF), les polychorobiphényls (PCB) 105, 114,118 et 123, le Polybromodiphénylether (PBDE) 183, le polybromobiphényl (PBB) 153 et plusieurs pesticides organochlorés incluant le trans-nonachlore, le cis-heptachlore-époxyde, dieldrine, le β hexachlorocyclohexane, et l'hexachlorobenzène. Les associations les plus fortes observées sont l'OCDF avec un OR (IC95%) ajusté sur l'âge et le BMI de 4,60 (2,47 ; 9,84) suivi de près par le cis-heptachlore-époxyde : 3,23 (1,81 ; 6,51). Les résultats de cette première analyse suggèrent aussi d'appliquer d'autres modèles statistiques plus pertinents pour mettre en évidence ces associations car ce sont des modèles « single pollutant » c'est-à-dire qu'ils n'évaluent que l'effet d'un polluant à la fois (9) (28). Une seconde analyse a donc été réalisée par le LABERCA en utilisant un autre modèle, basé aussi sur le principe de la régression logistique mais en appliquant un terme de pénalité sur les

coefficients : la régression logistique type «Elastic-net » (Section I.3), qui est une méthode de sélection de variables pour réduire la dimension des données. Cette méthode s'avère particulièrement adaptée pour des variables fortement corrélées. Par ailleurs, les coefficients retrouvés s'interprètent de la même manière qu'un odds ratio obtenu par une régression logistique standard. Les résultats montrent que l'OCDF, PCB 118, PCB 123, PBB 153, la dioxine 1.2.3.7.8 PeCDD, et plusieurs pesticides organochlorés comme le cis heptachlore époxyde sont associés avec le statut d'endométriose profonde (29). Au-delà de ces résultats, l'utilisation de ce modèle plus flexible de régression logistique est intéressante car on peut y inclure un grand nombre de co-variables. Néanmoins, cette méthode reste basée sur les principes d'une régression linéaire. Il serait intéressant de tester d'autres méthodes alternatives pour pouvoir aussi évaluer des combinaisons non linéaires, ce qui se rapprocherait plus d'une situation réelle (Voir les résultats des analyses préliminaires dans les tableaux A2 et A3 de l'annexe).

4. Limites des approches statistiques en épidémiologie environnementale

Traditionnellement, les approches statistiques utilisées dans le domaine de l'épidémiologie environnementale ont été développées autour de l'association entre l'exposition humaine aux composés individuels et la problématique de santé. Ainsi, l'effet de chaque substance est étudié séparément et indépendamment des autres. Une limitation majeure de cette approche est de ne pas tenir compte de la complexité des mélanges de produits chimiques et de leurs interactions potentielles, ce qui devrait être au contraire considéré pour s'approcher de scénarios d'exposition plus réalistes.
Aujourd'hui, grâce aux avancées technologiques d'analyse multi-résidus, nous avons accès à de plus en plus de profils d'exposition interne à plusieurs polluants chimiques, simultanément. Par conséquent, les études observationnelles sont capables de générer des matrices complexes d'exposition individuelles avec de nombreuses variables d'exposition pour chaque individu.
Si l'on considère un tel ensemble étendu de contaminants chimiques, ces différentes variables d'exposition peuvent être très redondantes et très corrélées entre elles, surtout si elles partagent les mêmes propriétés physicochimiques, sources d'exposition ou voies métaboliques, induisant une forte « multicollinéarité » des données. Or les méthodes statistiques les plus utilisées, comme les modèles de régression linéaire ou logistique (basées sur les moindres carrés), sont limitée peu adaptées pour les données très corrélées, car cela peut altérer la fiabilité des résultats et entraîner une instabilité importante des coefficients du modèle. De surcroît, ce sont des approches basées principalement sur l'étude des combinaisons linéaires entre variables.
Par conséquent, l'enjeu majeur de l'épidémiologie environnementale aujourd'hui, est de mieux comprendre la problématique des mélanges complexes en répondant à plusieurs questions parmi lesquelles comment identifier les composés au sein du mélange qui sont les plus associés à la problématique de santé d'intérêt, et quelles sont les interactions entre ces différents composés (30).

5. Intelligence artificielle, dont l'apprentissage automatique, en épidémiologie et en santé publique

a) Émergence de méthodes statistiques basées sur l'intelligence artificielle

L'utilisation de ces méthodes d'apprentissage automatique en épidémiologie et santé publique est en forte croissance. Comme dans d'autres domaines, les vastes bases de données de santé disponibles, nécessitent d'être valorisées et analysées alors qu'au dire de beaucoup de spécialistes elles sont actuellement sous-utilisées. Le traitement d'images médicales ou des séquençage génomique est une illustration frappante de disponibilité de données massives : il pose un défi dû à la fois à la complexité des données et au rapport inhabituel entre le nombre de variables et le nombre d'observations. Le nombre de variables est souvent considérablement plus grand que celui des observations : une image d'un mégapixels en couleur correspond à trois millions de variables… (31).
Aujourd'hui, de plus en plus d'articles sont consacrés à l'application de l'intelligence artificielle en santé : dans le moteur de recherche de PubMed, les articles contenant « machine learning » AND epidemiology », sont passés de 11 en 2010 à 382 en 2018.

b) Définition de l'apprentissage automatique

L'apprentissage automatique ou « machine learning » est une branche de l'intelligence artificielle, déjà utilisé dans de nombreux domaines (le marketing, l'économie, le monde du digital, …), par exemple, pour la détection de fraude par carte de crédit, les systèmes de recommandation et l'analyse des marchés financiers, les moteurs de recherche, ainsi que la reconnaissance faciale, et plus récemment, pour une nouvelle application telle que la prédiction de diagnostics médicaux (épidémiologie génétique, cancérologie, …).
L'objectif du « machine learning » est d'établir des structures à partir des données (ou profils) dans un but descriptif, explicatif, prédictif et/ou diagnostic. Ces méthodes visent à créer une règle minimisant le risque empirique d'erreur en prédiction, à partir d'exemples représentatifs d'une population. Elles sont particulièrement pertinentes et efficaces surtout lorsque le nombre de variables est largement supérieur aux nombres d'observations (ce que l'on appelle la multi dimensionnalité des données), là où les régressions logistiques et linéaires de base font défaut (32). Il existe deux principaux modes d'apprentissage automatique : l'apprentissage supervisé et le non-supervisé. L'apprentissage supervisé est guidé, dans l'objectif d'atteindre une tâche précise. C'est le mode d'apprentissage le plus utilisé avec un meilleur potentiel en épidémiologie, dont l'objectif est ici, d'établir une association entre les polluants et la problématique de santé. Pour cela, des données étiquetées lui seront données, c'est-à-dire que l'algorithme va savoir quelles sont les variables

explicatives (X, autrement dit, les polluants) et quelles sont les variables à expliquer (Y, autrement dit, la probabilité d'avoir la maladie). Tom Mitchell définit l'apprentissage supervisé comme : « Un programme informatique apprend lorsque sa performance à une tâche T, mesurée par P, s'améliore avec l'expérience E » (33). Dans le contexte de l'épidémiologie, la tâche ici serait de prédire la maladie (probabilité de développer la maladie), la performance : sa qualité de prédiction, et l'expérience E serait ici l'apprentissage sur les bases de données. Contrairement à l'apprentissage supervisé, le non-supervisé n'a pas de tâche précise, et les données ne sont pas étiquetées. L'algorithme va devoir établir des profils à partir de ses données brutes.

c) Applications de ces méthodes dans le domaine de la santé

Dans le rapport de Cédric Villani sur les stratégies nationales en matière d'intelligence artificielle et de santé, il est dit que celle-ci ouvre donc des perspectives très prometteuses pour améliorer la qualité des soins au bénéfice du patient et réduire leur coût à travers une prise en charge plus personnalisée et prédictive mais également leur sécurité grâce à un appui renforcé à la décision médicale et une meilleure traçabilité. Elle peut également contribuer à améliorer l'accès aux soins des citoyens, grâce à des dispositifs de pré diagnostic médical ou d'aide à l'orientation dans le parcours de soin. En effet, l'IA ouvre de nouvelles opportunités pour innover en construisant un diagnostic et une stratégie thérapeutique plus adaptés au besoin du patient, son environnement et son mode de vie. Elle permet en effet de mieux détecter les symptômes et de faire un suivi prédictif du développement d'une maladie et de son évolution ou pronostic, d'exploiter les résultats d'analyse (imagerie médicale…), de soumettre de nouvelles hypothèses de diagnostic et de formuler des propositions thérapeutiques plus personnalisées. Elles peuvent aussi améliorer la détection des effets secondaires d'un médicament lors des phases d'essais cliniques, et donc in fine avoir un impact positif sur l'innovation pharmacologique (meilleur ciblage thérapeutique, accélération et sécurisation de la mise des médicaments sur le marché, etc.). En matière de recherche médicale et d'épidémiologie, les technologies d'IA permettent de faciliter l'analyse de données massives, afin d'établir des profils, et de déterminer des facteurs de risque associés à une pathologie (34).

Beaucoup d'applications existent aujourd'hui dans la littérature, notamment dans le domaine de la cancérologie, neurologie, psychiatrie, cardiologie, et plus récemment d'autres domaine. Plusieurs exemples non exhaustifs : concernant l'aide à l'imagerie médicale dans différentes spécialités, nous pouvons citer comme exemple une étude visant à améliorer le diagnostic précoce de la maladie d'Alzheimer par une IRM cérébrale (35), ou alors une autre étude visant à établir un diagnostic de nodules ou de lésions du sein par échographie (36). Dilsizian et Siegel ont également discuté de l'application potentielle de l'intelligence artificielle pour donner un diagnostic personnalisé par l'imagerie cardiaque (37). D'autres applications existent aussi pour dépister des rétinopathies diabétiques sur des images de rétines (38). Concernant l'aide au diagnostic et à la thérapeutique sans

forcément d'imagerie : nous pouvons citer une étude récente utilisant l'apprentissage automatique pour analyser des photos de lésions de la peau afin repérer des mélanomes ou autres cancers de la peau (pour éviter la biopsie) (39). Une autre application récente a été étudiée afin de prédire la réponse thérapeutique des patients atteints d'un cancer du rectum avancé traités par radio chimiothérapie préopératoire (40) (41) (42).
Ces méthodes d'apprentissage automatique ont démontré un fort potentiel de capacité prédictive mais l'intérêt et l'objectif ici, dans le cadre de l'épidémiologie (notamment environnementale), n'est pas de prédire la pathologie, mais de mieux connaître et d'appréhender les structures de données complexes, très corrélées entre elles, et d'établir un lien entre les expositions aux polluants et une pathologie. Elles permettent de s'affranchir de certaines limites et problématiques citées ci-dessus. Néanmoins la performance et l'applicabilité de ces méthodes afin d'établir des liens de causalité restent encore peu explorées.

6. Revue de la littérature des méthodes statistiques d'intérêt pour aborder l'effet « cocktail » en épidémiologie environnementale

Au cours de la dernière décennie, l'émergence de nouvelles méthodes d'analyse des données en statistique et en informatique dans le domaine de l'apprentissage automatique a élargi le spectre des algorithmes disponibles tout en prenant en compte les contraintes associées à ce type de données. Un exemple de cet intérêt grandissant pour ces méthodes, est le Workshop « Statistical Approaches for Assessing Health Effects of Environmental Chemical Mixtures in Epidemiology Studies » de l'Institut National des Sciences de la Santé Environnement (NIEHS) réalisé en 2015 pour étudier et comparer ces méthodes avancées en épidémiologie environnementale (43).
Sur la base d'une revue de littérature préliminaire, plusieurs études ont considéré et comparé différentes méthodes alternatives aux approches de régression linéaire ou logistique, traditionnellement appliquées en épidémiologie environnementale. La plupart de ces études traitent notamment de la pollution de l'air comme source d'exposition, dont une grande partie utilise des données simulées (44) (30) (43). Une autre partie des études relèvent du domaine de l'épidémiologie génétique car leurs bases de données sont maintenant très conséquentes (45).
Ces études classent généralement les méthodes d'apprentissage (supervisées ou non-supervisées) en trois groupes : les méthodes de sélection de variables, de réduction de la dimension et les méthodes de classification et de prédiction.
Parmi les méthodes de **sélection des variables**, qui « neutralisent » les variables non pertinentes et qui identifient le meilleur sous-ensemble de variables les plus associées à la pathologie, on y retrouve les régressions pénalisées ou méthodes de régularisation, souvent utilisées en épidémiologie environnementale comme alternative aux modèles classiques : Least Absolute Shrinkage and Selection Operator (LASSO), Elastic-Net, régression en arête (Ridge),…, dont le but est d'appliquer

un terme de pénalité pour lisser les coefficients non pertinents jusqu'à zéro ou proche de zéro (46) (47). Une autre méthode de sélection de variables est l'algorithme de délétion/substitution/addition (DSA), qui est un modèle de régression itératif enlevant un terme, puis le substituant et ajoutant un autre terme. Celui-ci est moins fréquemment utilisé en épidémiologie environnementale (48).

Concernant les approches **de réduction de la dimension**, les méthodes les plus connues sont l'Analyse en composantes Principales (ACP) et la régression des moindres carrés partiels (régression PLS), qui sont beaucoup utilisées notamment dans le domaine de la chimio métrie et commencent à prendre de l'ampleur dans le domaine de l'épidémiologie environnementale. L'ACP est une méthode non-supervisée tandis que la PLS est une méthode supervisée. Elles sont toutes deux principalement fondées sur des combinaisons linéaires (49) (50).

Enfin dans le groupe des **méthodes de discrimination** ou **de prédiction**, les plus connues sont les réseaux de neurones, les arbres décisionnels et les arbres « boostés », et les machines à vecteurs de support. Ce sont des méthodes largement utilisées en épidémiologie génétique mais très peu en épidémiologie environnementale. Ces méthodes semblent très performantes lorsque les données sont très corrélées mais aussi lorsque le nombre de variables est supérieur au nombre d'observation. De plus, elles permettent d'étudier des relations non linéaires entre les variables ce que ne fait pas les modèles des moindres carrés. Cependant, malgré leur grande performance prédictive, elles apparaissent souvent comme des « boîtes noires » à cause de difficultés à interpréter leurs sorties qui ne sont pas similaires avec celles que les épidémiologistes ont l'habitude de traiter, mais aussi à cause de l'incompréhension des processus complexes impliqués dans ces algorithmes (51) (52) (53). Un tableau récapitulatif de ces quelques méthodes avec leurs forces et faiblesses, ainsi que les articles les ayant appliquées est disponible en Annexe (Tableau A1).

7. Objectifs

Dans ce contexte, notre présente étude a pour ambition de sélectionner puis d'appliquer différentes méthodes issues de l'apprentissage statistique, pour mettre en évidence une association entre les niveaux internes en polluants organiques persistants (biomarqueurs d'exposition) et la présence d'endométriose.

Ce travail a fait appel à une première partie de revue de littérature pour pouvoir justifier la sélection des différentes méthodes. Celles-ci ont ensuite été mises en œuvre afin d'identifier un ensemble de marqueurs d'exposition pertinents au regard du lien avec la pathologie considérée. L'émergence récente de ces nouvelles approches en épidémiologie explique probablement qu'elles soient à ce jour principalement mises en œuvre dans des exercices de simulation. Le travail entrepris vise à les éprouver dans le cadre d'une application réelle, et ainsi d'évaluer leur performance et leur utilité à partir d'un jeu de données expérimental relatif à l'étude du lien entre les niveaux exposition chimique interne mesurés chez des femmes adultes et l'endométriose. Un objectif secondaire a été de comparer

ces modèles entre eux ainsi qu'à un modèle plus classique, en termes de résultats mais également en termes de conditions et facilité de mise en œuvre.
Ce travail a été conduit en collaboration avec le laboratoire d'étude des résidus et contaminants dans les aliments (UMR 1329 INRA LABERCA) et l'unité Statistique, Sensométrie et Chimio métrie (StatSC), deux unités de recherche de l'école nationale vétérinaire, agroalimentaire et de l'alimentation Nantes atlantique (Oniris).

V. MATERIEL

1. Présentation de la base de données : projet ENDOTOX

Nous avons basé notre travail sur une base de données relative aux concentrations de polluants mesurés dans le tissu adipeux des sujets inclus dans notre étude clinique support, car d'une part leurs taux de détection sont plus élevés que dans le sérum et d'autre part, davantage d'associations significatives ont été révélées dans une première approche à partir de ces marqueurs tissulaires considérés individuellement. Dans le Tableau 1 sont présentées les grandes familles de polluants retrouvés dans le tissu adipeux des cas et des témoins. Une liste exhaustive de ces polluants est disponible dans le Tableau A4 en annexe.
Pour cette étude, nous avons sélectionné seulement 2 facteurs de confusion connus comme étant associés aux niveaux d'exposition interne aux polluants organiques persistants, à savoir l'âge et l'IMC (Tableau 2). Des analyses exploratoires préliminaires ayant montré le caractère non déterminant d'un troisième éventuel facteur de confusion qu'est la durée d'allaitement, ce dernier n'a pas été inclus comme co-variable dans nos modèles. Ceux-ci peuvent aussi être une conséquence de l'endométriose. Enfin, le statut d'endométriose profonde invasive (oui/non) est considéré comme la variable binaire à expliquer.

Tableau 1. Polluants (variables explicatives) recueillis chez les cas et les témoins

Famille de Polluant	Nombre de marqueurs d'exposition
Dioxines	17
Polychlorobiphényles (PCB) dioxine-like + PCB non dioxine-like	12 + 6
Polybromodiphényléthers (PBDE)	8
Polybromobiphényles (PBB)	1
Hexabromocyclododécane (HBCD)	1
Pesticides organochlorés (OC)	8
TOTAL	53

Tableau 2. Démographie de la population des cas/témoins (Facteurs associés)

	Control	Cases	P-value
Age (années)	32,6 (±6,5)	34,3 (±6,2)	0,13
IMC (kg/m²)	25,4 (±5,9)	24,0 (±5,1)	0,22

2. Logiciels

Pour réaliser les analyses statistiques, nous avons utilisé le logiciel R version 3.4.3 ainsi que le logiciel R Studio version 1.1.383 et le logiciel Microsoft Excel (2007) pour gérer la base de données. Les packages R utilisés sont listés dans le Tableau A5 présenté en annexe.
Concernant la revue de littérature, nous avons utilisé le logiciel Zotero (version 5.0.74) pour gérer et référencer les articles choisis. Nous nous sommes appuyés uniquement sur le site PubMed pour appliquer la chaîne de recherche et identifier les articles.

VI. MÉTHODES

1. Sélection des modèles

Une première étape de revue de la littérature a été réalisée pour sélectionner et justifier le choix des modèles. Les résultats de celle-ci ont été introduits dans la partie contexte Section I.3. L'objectif était de recenser les articles ayant appliqué ces méthodes alternatives aux méthodes classiques dans le domaine de l'épidémiologie. Pour réaliser cette revue de la littérature, nous avons établi une chaîne de recherche avec mots clés, contenant 2 parties : une première relative à la modélisation (« machine learning », « statistic approach », « multipollutant model », et une seconde relative à l'épidémiologie et l'environnement (« health », « environment », « epidemiology »,). En appliquant cette chaîne sur PUBMED, nous avons obtenu 1691 articles et nous n'en retenons que 21 (voir le tableau A2 de l'annexe) en gardant principalement les articles focalisés sur l'. Nous nous sommes également appuyés sur plusieurs articles comparant les différents modèles alternatifs existants, notamment le workshop de NIEHS (43), qui conclut que la performance des méthodes est données dépendante et qu'aucune option définitive ne peut être privilégiée. Sur la base de cette revue, nous avons finalement retenu 3 approches méthodologiques pour évaluation et comparaison dans le cadre de notre application particulière :

- Les réseaux de neurones artificiels (Artificial neural network : ANN)
- Les machines à vecteurs de support (Support Vector Machines, SVM)
- Les « Boosting trees »

Les critères de sélection de ces 3 méthodes sont les suivants : Premièrement, le fait qu'elles soient très peu utilisées dans le domaine de l'épidémiologie environnementale mais qu'elles prennent beaucoup d'ampleur en épidémiologie génétique ainsi que dans les études de prédiction de cancer. Cela constitue en partie l'originalité de notre sujet. Par ailleurs, ces méthodes s'affranchissent des contraintes des modèles classiques : elles sont notamment adaptées à des données où le nombre de variables (p) est supérieur au nombre d'observation (n), et elles peuvent intégrer de nombreuses covariable très corrélées et évaluer l'association de plusieurs polluants et non indépendamment des uns des autres. Elles peuvent également traiter des combinaisons non linéaires entre les variables. Nous allons utiliser ici la régression logistique pénalisée Elastic-Net, comme modèle de référence pour comparer ces algorithmes en se référant à l'analyse préliminaire effectuée par le LABERCA, car ce modèle est celui qui est à la fois le plus proche de la régression logistique simple en termes d'interprétation, mais à la fois d'avantage comparable aux trois autres modèles.

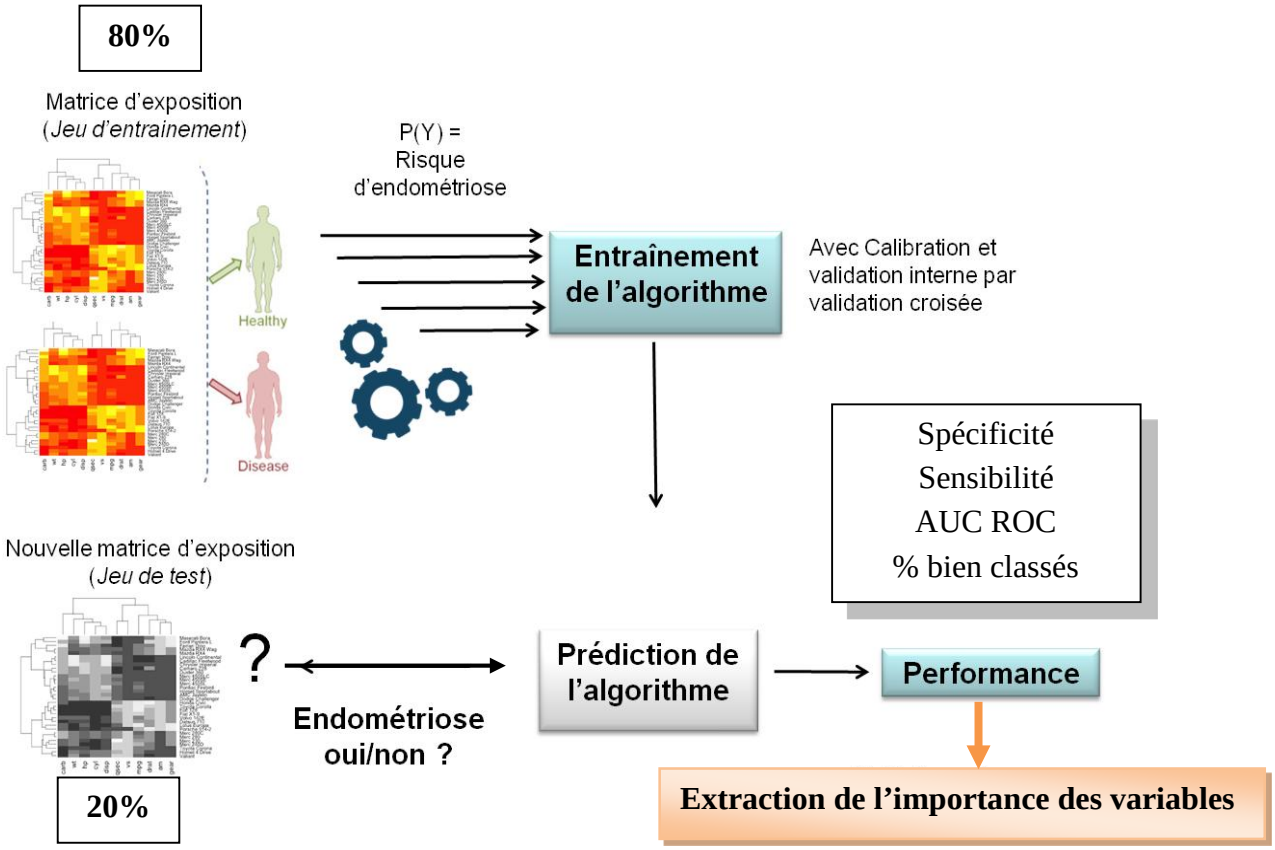
2. Schéma de modélisation et d'application des algorithmes

Pour les trois modèles nous avons suivi un schéma de modélisation contenant les étapes suivantes illustrées dans la figure 1 :

a. Préparation des données
b. Séparation du jeu de données en un jeu d'entraînement et jeu de test
c. Entraînement des modèles et calibration des paramètres
d. Ajustement et évaluation des différents modèles sur le jeu de test

Toutes les étapes de modélisation ont été réalisées grâce au package « caret » du logiciel R, qui est un package regroupant un ensemble de fonctions permettant de rationnaliser le processus de création de modèles prédictifs. L'avantage de ce package est qu'il est simple à utiliser et les fonctions de validation croisée ou de bootstrapping y sont incluses.

Figure 1. Illustration du schéma de modélisation et d'application des algorithmes d'apprentissage



a) Préparation des données

Il est important de préparer les données en amont, avant d'entraîner l'algorithme. Ici, dans nos données, il y a beaucoup de concentrations variées de polluants, avec différentes unités. Une première étape consiste à traiter les données manquantes. Pour ce faire, nous avons choisi une méthode d'imputation des données avec un algorithme spécifique (« mice function » du package « mice » dans R) c'est-à-dire que chaque variable incomplète va être imputée par un modèle séparé. La deuxième étape va commencer à uniformiser les variables « polluants » en appliquant la fonction Logarithme sur les concentrations de polluants. Enfin, la dernière étape de préparation est de centrer-réduire les variables « polluants ». Cela consiste à modifier les variables de façon à fixer leur moyenne (à 0) et leur variance (à 1), ce qui équivaut à un changement d'échelle pour mieux uniformiser les données et pour pouvoir les comparer.

b) Séparation du jeu de données en un jeu d'entraînement et jeu de test

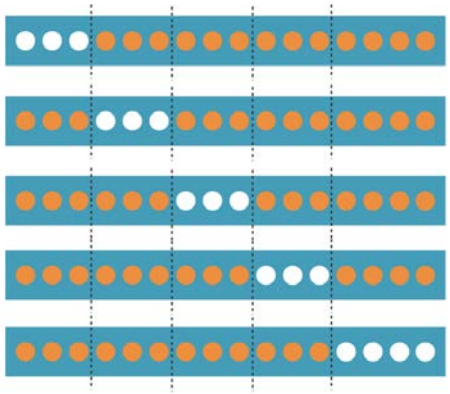
Cette étape est nécessaire pour l'application d'un modèle d'apprentissage automatique : on va donc séparer le jeu de données en 2 jeux de données différents. Le premier jeu sera appelé « le jeu d'entraînement » (training set) et nous avons choisi d'inclure 80 % des observations dans ce jeu. Ce jeu, comme son nom l'indique va servir l'algorithme à s'entraîner et rectifier ses erreurs, pour mieux connaître le profil d'exposition des femmes ayant une endométriose et celles n'en ayant pas.
Le deuxième jeu sera le « jeu de test » ou « jeu de validation » et il contiendra les 20 % restant des observations. Celui-ci va juste contenir les variables d'exposition, et l'algorithme devra prédire si la femme a une endométriose ou non. Les données de ce jeu de test n'auront jamais servi à l'entraînement, elles servent uniquement pour la prédiction. Une fois que l'algorithme a établi sa prédiction, nous allons comparer avec la variable endométriose sur nos données de test pour évaluer la capacité de prédiction de l'algorithme.
Concernant la justification du pourcentage de séparation : il est recommandé dans la littérature d'avoir plus de données pour permettre à la machine d'apprendre, que pour tester et prédire, et dans la plupart des cas on retrouve cette proportion 80%, 20%.

c) Entraînement des modèles et calibration des paramètres

Dans l'étape d'entraînement, l'algorithme va apprendre des données sur le jeu d'entraînement et va essayer d'établir les profils d'exposition des cas et des témoins. Mais pour appliquer un algorithme, des paramètres doivent être choisis au préalable : chaque méthode possède différents paramètres (expliqués ci-après). Cette étape va permettre de choisir sont les meilleures valeurs des paramètres.

Elle va permettre aussi de pouvoir évaluer et comparer les différents packages du logiciel R qui n'appliquent pas forcément les mêmes algorithmes pour une même méthode.
Pour optimiser cette étape, nous allons utiliser ici une méthode de rééchantillonnage qui est la validation croisée (k-fold cross validation). Cette méthode consiste à diviser le jeu d'apprentissage en k parties, une seule des k parties va servir à la validation interne de l'algorithme et le reste va servir à l'apprentissage proprement-dit. Cette opération va se répéter k fois afin que chaque sous-échantillon puisse servir une fois pour la validation interne (Figure 2).
Par ailleurs, cette validation croisée sera stratifiée sur la variable endométriose, c'est-à-dire que l'on va répartir équitablement le nombre de cas d'endométriose et de témoins dans chaque partie.
Dans la revue de littérature, la valeur de cas est souvent de 5 ou 10. Ici, étant donnée la petite taille de notre échantillon, nous choisirons une validation croisée en 5 parties.

Figure 2. Exemple d'une validation croisée en 5 parties sur un jeu de données de 16 observations. Chaque point appartient, tour à tour, à une des 5 parties pour la validation interne (en blanc) ou aux 4 autres parties pour l'entraînement (en orange). (Red. « Évaluez vos modèles sans sur-apprentissage » ; OpenClassrooms.com)



d) Ajustement et évaluation des différents modèles sur le jeu de test

Nous allons ensuite évaluer l'algorithme optimisé avec de nouvelles observations d'exposition sur lesquelles il n'aura pas appris (le jeu test). Il va ainsi, pour des valeurs fixées des paramètres de l'algorithme, prédire la probabilité d'endométriose pour chaque nouvelle observation. Puisque dans ce jeu test nous connaissons les femmes ayant l'endométriose et celles ne l'ayant pas, nous allons comparer ces données avec ce que l'algorithme aura prédit, et nous allons ainsi mesurer sa performance de prédiction.

Pour mesurer sa performance, nous avons choisi plusieurs indicateurs :

- La sensibilité : qui correspond ici, à la probabilité que l'endométriose soit prédite par l'algorithme quand il s'agit réellement de femmes atteintes de l'endométriose,
- La spécificité : correspond ici, à la probabilité de déclarer qu'il ne s'agit pas d'un cas d'endométriose quand il s'agit de témoins,
- Le pourcentage de bien classés, c'est-à-dire le % d'individus que l'algorithme a prédit correctement (par rapport aux cas réels). Il combine la sensibilité et la spécificité,
- L'aire sous la courbe ROC (AUC) : la courbe ROC est une courbe sensibilité/spécificité. Généralement, cette courbe donne le taux de vrais positifs en fonction du taux de faux positifs. L'aire sous la courbe permet donc d'avoir un aperçu du taux de distinction par l'algorithme d'une personne témoin (non atteinte de l'endométriose) et d'un cas (personne atteinte de l'endométriose)

Hormis la performance de prédiction de ces modèles, ce que nous recherchons principalement ici est de savoir quelles sont les variables les plus pertinentes et les plus associées à l'endométriose. Car si un algorithme est performant en termes de prédiction, c'est qu'il est capable d'établir le profil d'exposition d'une personne qui a l'endométriose comparé au profil d'un témoin. Pour ce faire, nous allons extraire via ces algorithmes, le classement des variables par ordre d'association à l'endométriose que l'on nommera ici l'importance des variables.

3. Description et calibration des modèles sélectionnés

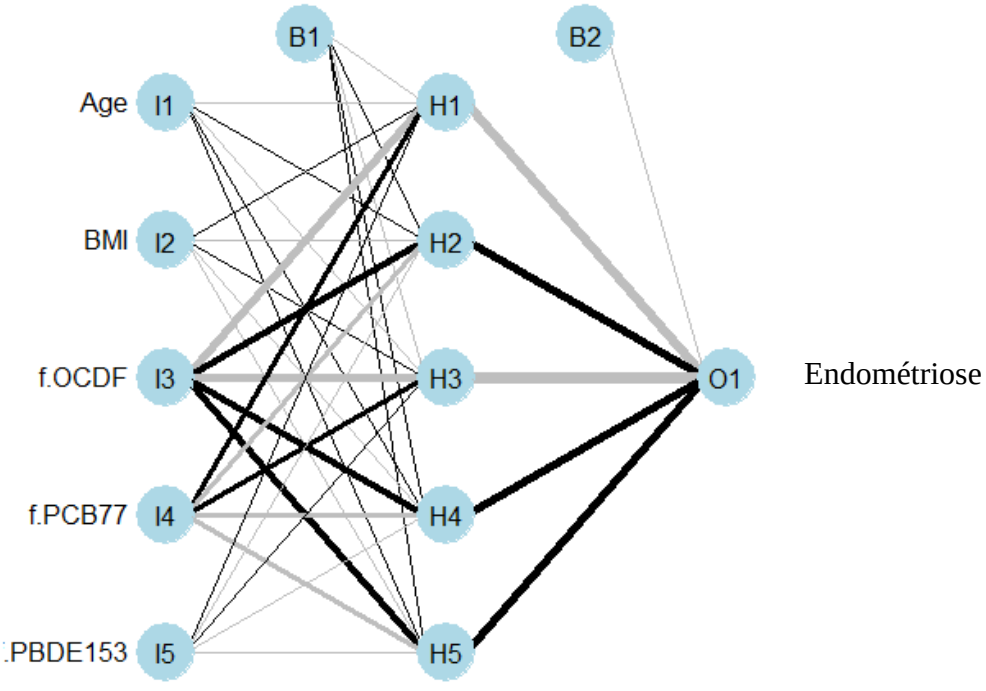
a) Réseaux de neurones artificiels

i. Description du modèle

Depuis quelques années, l'utilisation des réseaux de neurones artificiels (ANN) dans la prédiction de pathologies ou dans la classification de pathologies, est croissante. Les ANN sont des algorithmes d'apprentissage automatique qui s'inspirent des réseaux de neurones biologiques.
Le modèle de ANN le plus utilisé et le plus simple est le perceptron multi-couches : il comporte 1 couche d'entrée avec un ou plusieurs neurones représentant ici les variables explicatives et les facteurs associés, 1 ou plusieurs couche(s) cachée(s) avec un ou plusieurs neurones représentant des variables latentes, et 1 couche de sortie avec un (ou plusieurs) neurones de sortie représentant la variable réponse ici la présence d'endométriose (Figure 3). Chaque neurone est relié avec les autres neurones de la couche précédente par des synapses sur lesquelles est attribué un poids : le poids synaptique.

Il existe aussi les biais : un biais est un neurone dans lequel la fonction d'activation est en permanence égale à 1. Comme pour les autres neurones, un biais se connecte aux neurones de la couche précédente par l'intermédiaire d'un poids synaptique.
Les perceptrons multi-couches sont des ANN non bouclés, c'est-à-dire que l'information ne circule que dans un seul sens, celui de l'entrée vers la sortie. Ils sont capables de traiter les phénomènes non linéaires.
L'apprentissage de ce modèle va consister à déterminer ces poids synaptiques en minimisant l'erreur entre les sorties désirées et les valeurs observées.
Pour appliquer ce type de ANN, il est nécessaire de choisir au préalable, différents paramètres qui dépendent de la taille de l'échantillon mais aussi des différents packages du logiciel statistique R.

Figure 3. Réseau de neurones, perceptron multi-couche réalisé avec la base de données ENDOTOX. Légende : I = couche/neurones d'entrée (variables explicatives + facteurs associés) ; H=couche/neurones cachée (variables latentes) ; O=couche/neurones de sortie (Probabilité d'endométriose) ; B1, B2 = neurones Biais



ii. Calibration du modèle

➢ Paramètres à définir

Plusieurs paramètres sont à définir au préalable :

- Le nombre de couches cachées

Le Perceptron multicouche (PMC) à une seule couche cachée est un approximateur universel capable de modéliser n'importe quelle fonction continue (54). Par ailleurs, dans la revue de littérature, les études épidémiologiques ayant appliqué les ANN, utilisent pour la plupart une seule couche cachée. Cependant, depuis l'arrivée de l'apprentissage profond (« deep learning »), et pour des données très volumineuses et complexes, il serait intéressant d'utiliser plusieurs couches cachées, mais il n'existe actuellement pas de réponse arrêtée quant à leurs nombres.

- Nombre de neurones cachés

Parallèlement au nombre de couches cachées, il n'existe pas non plus de consensus sur le nombre de neurones par couches cachées, mais deux formules ont été suggérées dans la littérature pour les estimer (55). La première formule concerne le nombre de neurones dans la première couche cachée (Formule 1) et la deuxième formule concerne le nombre de neurones dans la 2ème couche cachée (Formule 2) :

(Formule 1) $$x1=\sqrt{((m+2)\times N)}+2\times\sqrt{(N/(m+2))}$$

(Formule 2) $$x2=m\times\sqrt{(N/(m+2))}$$

où N = nombre de variables d'entrées = 53 ici (nombre de polluants), and m = nombre de variable de sortie = 1 ici (endométriose oui/non). Soit x1=21.4 et x2= 4.28. D'autres suggèrent aussi un nombre de neurones cachés entre m et N (soit entre 1 et 53), ou alors 75%N+m, soit 41 ici (56).

Dans notre étude, nous avons testé des modèles pour voir l'effet de différents nombres de neurones de la couche cachée à l'aide des mêmes indicateurs que ceux utilisés ci-dessus, à savoir la sensibilité, la spécificité, et l'AUC.

- Paramètre de régularisation (decay) :

Ce paramètre est une pénalité appliquée sur les poids synaptiques extrêmes pour éviter le sur-apprentissage.

La plupart des études épidémiologiques utilisant ce paramètre entre 0.001 et 1.

Pour le déterminer, nous allons tester la performance des différents modèles avec différentes pénalités.

- Packages R

Pour cette étude, nous avons testé le package « nnet » (57) à l'aide du package « caret » (58). Ce package ne nous permet pas de choisir le nombre de couches cachées car le paramètre est fixé à 1, mais nous pouvons faire varier le nombre de neurones cachés et le paramètre de régularisation (« weight decay ») pour éviter le sur-apprentissage. Nous avons également testé le package « neuralnet » (59) qui pouvait quant à lui faire varier le nombre de couches cachées.

iii. Estimation de l'importance des variables

Au niveau de l'interprétation, on sait que pour un réseau de neurones sans couche cachée (H=0), le poids synaptique reliant la couche d'entrée à la couche de sortie, est équivalent aux coefficients obtenus à l'aide d'une régression logistique.
Cependant à partir d'une couche cachée, les poids ne peuvent plus s'interpréter de la même manière qu'une régression logistique car l'on ajoute des niveaux de complexité. Il existe dans la littérature plusieurs formules utilisant les poids synaptiques pour déterminer l'importance relative des variables d'exposition (polluants et facteurs de confusion). Les deux principales sont :
- La formule d'Olden, appelée la connexion des poids, c'est-à-dire que le poids attribué à un polluant pour l'endométriose est égal à la somme des produits des poids reliant ce polluant aux neurones de la couche cachée puis de ces neurones de la couche cachée à la couche de sortie qu'est l'endométriose (60).
- La formule de Garson qui sépare les poids synaptiques de la couche cachée à la couche de sortie en composants associés avec chaque neurone d'entrée (polluants) utilisant seulement la valeur absolue des poids synaptiques (60).
Dans une étude de simulation qui compare les différentes méthodes de classement de l'importance des variables par rapport à la variable réponse, la formule la plus performante est la première (en utilisant la fonction olden du package « NeuralNetTools » dans R), car elle identifie correctement la vraie importance de toutes les variables dans le réseau y compris celle avec de faible corrélation avec la variable réponse (61) (60).

b) Support Vector Machine

i. Description

Les machines à vecteurs de support (ou « support vector machine » en anglais : SVM) sont des modèles d'apprentissage supervisé utilisés initialement et principalement pour des problèmes de classification binaire. Le fonctionnement principal des SVM est d'établir un hyperplan permettant de séparer au mieux les groupes à prédire. Ici avec nos données, il s'agira de séparer le groupe témoin

(pas d'endométriose) du groupe cas (endométriose). Il existe plusieurs algorithmes de SVM, un pour les données linéairement séparables et plusieurs pour celles non linéairement séparables.

Concernant les données linéairement séparables (Figure 4), il peut exister plusieurs hyperplans permettant de séparer correctement les 2 groupes. L'objectif de l'algorithme va être de trouver l'hyperplan optimal en maximisant la distance entre celui-ci et les données d'apprentissage. Les points les plus proches de cet hyperplan sont appelés les vecteurs supports. Pour classifier les nouvelles données, la machine va simplement regarder de quels côtés de l'hyperplan elles se situent pour déterminer à quel groupe elles appartiennent.

Concernant les données non linéairement séparables, l'algorithme va utiliser une fonction appelée « noyau » (kernel trick) qui va permettre de transformer les données (à l'aide d'un produit scalaire) pour les représenter dans un espace de plus grande dimension afin de trouver un hyperplan séparateur (Figure 5).

L'idée fondamentale derrière l'algorithme des SVM est de représenter les données dans un nouvel espace où elles sont, plus ou moins, linéairement séparables et de déterminer une règle linéaire (ou hyperplan) permettant de séparer au mieux les groupes. Plusieurs fonctions de noyaux sont disponibles : le noyau gaussien, le plus connu, ou encore polynomial (62).

Nous allons nous intéresser ici au SVM avec noyau gaussien, i.e. à fonction de base radiale, car au vu de la littérature, c'est l'algorithme le plus utilisé des SVM pour des relations non linéaires.

Figure 4. Graphique représentant en gris une classe (qui pourrait être ici les témoins) et en bleu les cas. Une multitude d'hyperplans peuvent permettre de séparer parfaitement les deux classes (Fig. 4.a). L'optimisation avec la marge permet d'obtenir un hyperplan le plus éloigné des 2 classes (Fig 4.b). Pour éviter le sur-ajustement, cette marge peut être modifiée en autorisant un faible taux de mauvaises classifications (Fig 4.c) (source : Bzdok, Krzywinski & Altman, 2018 (63)).

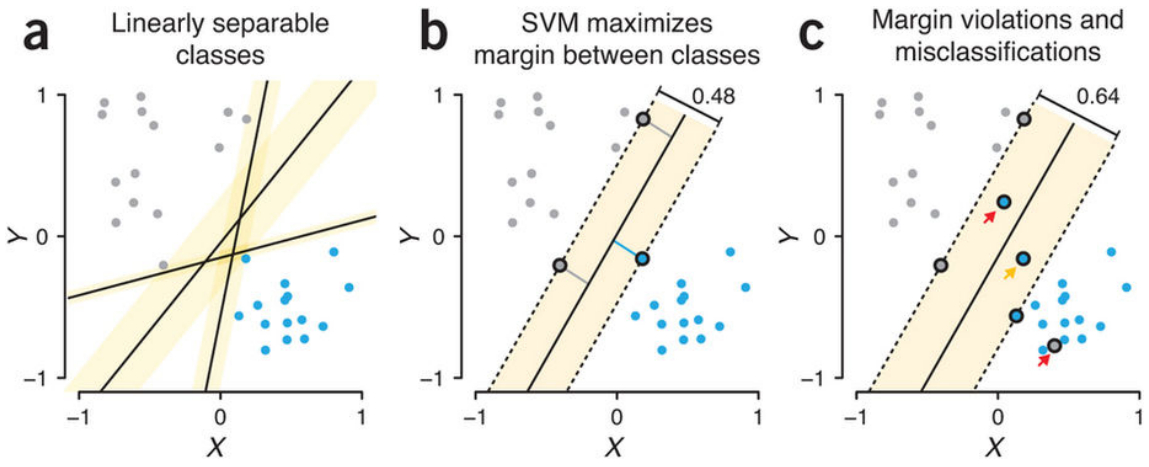
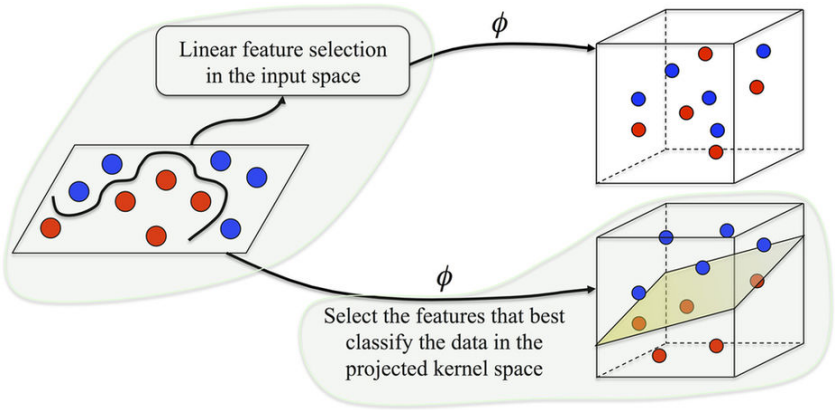


Figure 5. Méthode du noyau pour les données non linéairement séparables Ce graphique illustre la représentation des données dans un espace de dimension supérieure afin de pouvoir les séparer linéairement. (source : Adeli et al., 2017 (64)).



ii. Calibration du modèle

➢ Paramètre à définir :

Pour les SVM, plusieurs paramètres sont à définir :

- Le paramètre C (fonction de coût) :
  Il correspond au compromis entre l'erreur de classement, et la largeur de la marge. Plus ce paramètre est grand, moins il y a d'erreurs de classement mais plus il y a de risque de sur-apprentissage, ce qui sous-entend que la variance est faible mais les biais sont élevés. Au contraire, plus ce paramètre est petit, plus on augmente le risque d'erreur de classement mais on diminue le risque de sur-apprentissage.
- Le paramètre sigma ou gamma :
  Ce paramètre est à déterminer uniquement lorsque l'on ajoute la technique du noyau à la SVM. Il correspond à la largeur du noyau.

Le guide de Hsu, Chang et Lin (2003) (65), est souvent cité pour l'aide à l'application et à la calibration des SVM. Il suggère de tester l'algorithme avec, par exemple, le paramètre C compris entre 2.e-5 et 2. e+15 et le paramètre sigma compris entre 2.e-15 et 2. e+3, puis d'extraire le meilleur modèle avec le meilleur paramètre

Nous allons tester ici la performance prédictive de plusieurs modèles avec différents paramètres : soit en choisissant un algorithme et en testant une fourchette des différents paramètres, soit en fixant en avance les paramètres et en testant la performance de différents algorithmes. Nous avons testé, par exemple, un premier modèle en donnant une fourchette du paramètre C entre 1.e-4 à 1. e+4 et du

paramètre sigma du noyau entre 1.e-4 et 1. e+4, puis nous avons testé différents modèles avec des paramètres fixés (par exemple un modèle avec C = 1 et sigma = 0.01).

- Packages R

Pour les SVM, nous avons testé la fonction la plus communément utilisée : « svm » du package « e1071 » toujours via le package « caret » (66).

iii. Estimation de l'importance des variables

En ce qui concerne les SVM, il n'existe pas de formules ou de mesures spécifiques comme les formules d'Olden ou de Garson dans le cas des réseaux de neurones. Ce qui est pris en compte dans les SVM pour classer l'importance des variables est l'AUC de chaque variable explicative prise séparément. On va donc analyser la qualité prédictive de chaque variable (chaque polluant, l'âge et l'IMC), et nous allons les classer en fonction de leur AUC. Plus une variable aura un indice AUC proche de 1, mieux elle sera classée. Comme avec les réseaux de neurones, nous avons échelonné ici l'importance des variables entre 0 et 100.

c) Boosting trees

i. Description

- Arbres de segmentation (Figure 6) :

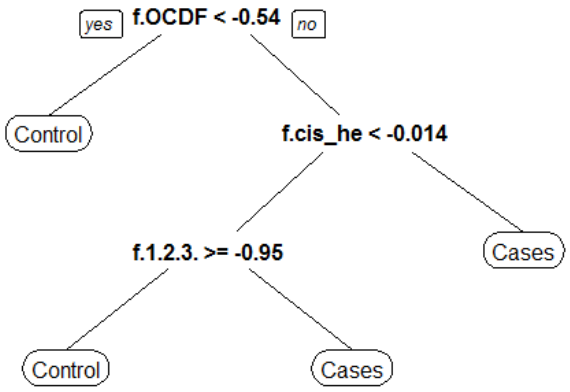
Ce sont des modèles de discrimination non-paramétriques. Ils sont constitués d'une suite de séparations binaires qui visent à séparer un ensemble d'individus en 2 sous-ensembles ayant chacun des caractéristiques communes en lien avec la variable d'intérêt (ici l'endométriose). Ainsi l'objectif est, à chaque division binaire, de définir des sous-ensembles qui soient les plus homogènes possibles et ne contienne quasiment que des individus d'une seule catégorie (cas ou témoin).
Cette structure prend la forme d'un arbre, où chaque nœud indique quelle est la variable explicative choisie pour la séparation en deux sous-ensembles, prise parmi les polluants, l'âge et l'IMC. Les feuilles (nœuds terminaux) fournissent une segmentation la variable endométriose en cas ou témoin. Une interprétation plus détaillée d'un arbre de segmentation se trouve dans l'annexe section VIII.6.

- Le Boosting :

Le Boosting, est une technique de « méta-classifieurs », c'est-à-dire une méthode ensembliste applicables à de très nombreux algorithmes. Concrètement un modèle de base, souvent simple, va être appliqué de manière itérative pour optimiser la qualité de la prédiction recherchée. Ici, l'algorithme de base est celui des arbres de segmentation. C'est pour cet algorithme que la technique de boosting

a principalement été utilisée jusqu'ici. Le principe du boosting est de construire, ici, une famille d'arbres de segmentation, chaque modèle étant une version adaptative du modèle précédent, en donnant plus de poids aux observations mal classées. C'est le principe des « apprenants faibles » qui consiste à combiner les apprenants faibles (ceux dont les prédictions sont à peine peu supérieures à un choix aléatoire) pour obtenir un « apprenant fort ». Le résultat final est une somme pondérée de tous les classifieurs boostés.

Figure 6. Exemple d'arbre de segmentation (CART) appliqué à la base de données ENDOTOX



Il existe de nombreux algorithmes de boosting différents qui sont des variantes basées sur le principe expliqué ci-dessus. Le plus populaire étant celui nommé « AdaBoost.M1 » (31) implémenté dans le package R « fastAdaboost » par exemple.

ii. Calibration du modèle

- Paramètres à définir

S'agissant de l'algorithme « AdaBoost.M1 », un seul paramètre est à définir : celui du nombre d'arbres de segmentation que l'on veut prendre en compte, c'est-à-dire le nombre d'itérations. Il n'existe pas non plus dans la littérature de consensus concernant la valeur que devrait prendre ce paramètre. Généralement, on teste le modèle entre 1 itération et 1000 itérations. Les arbres de segmentation sur lesquels l'algorithme « AdaBoost.M1 » va appliquer chaque itération, sont des arbres simples, ne comprenant qu'un seul nœud, donc une seule séparation.

➢ Packages R :

Nous allons utiliser le package « fastAdaboost » reconnu dans la littérature (67), et toujours via le package « caret ».

iii. Estimation de l'importance des variables

Le classement de l'importance des variables les plus reliées à l'endométriose peut se faire grâce à une fonction du logiciel R. Pour classer ces variables la formule utilisée est une réduction de la fonction perte (la moyenne des carrés des erreurs) pour chaque variable et pour chaque séparation au niveau des arbres.

VII. RESULTATS

1. Réseaux de neurones artificiels

a) Optimisation et choix des paramètres

➢ Effet du nombre de couches cachées

Nous avons testé avec le package « neuralnet » différents nombres de couches cachées allant de 1 à 50. Le meilleur résultat été observé pour une et 2 couches cachées. Pour cette étude nous avons décidé de choisir un réseau de neurones avec une seule couche cachée, car plus il y a de couches, plus le modèle est complexe et donc difficile à interpréter car l'on obtient une multitude de poids à analyser. De plus, le modèle « nnet » à une seule couche avec pénalité est performant en termes de prédiction dans nos résultats.

➢ Effet du nombre de neurones cachés

Nous avons testé plusieurs modèles avec un nombre de neurones cachés ascendant allant de 5 à 50 (Section III.3.a.i). Dans les résultats de performance (Tableau 3), on constate que tous les modèles ont la même performance sauf pour le modèle contenant 10 neurones cachés. La performance des modèles avec 5, 25, 35, 45, 50 neurones caché est 88% de sensibilité du modèle, 64% de spécificité, 74% de réussite de classement, et une aire sous la courbe ROC de 76%. Le modèle à 10 neurones cachés, a un taux de réussite plus faible de 68% et une aire sous la courbe de 69%.

➢ Effet du paramètre de régularisation :

Il n'y a pas vraiment d'effet croissant ni décroissant de la performance avec l'augmentation ou la diminution de ce paramètre, il est aussi performant avec une valeur de 0 qu'une valeur de 0.9, en revanche, il est moins performant à 0.5.

Tableau 3. Résultats de la performance de prédiction des ANN avec calibration des paramètres. Abréviations : CC = nombre de Couches Cachées ; NC = nombre de Neurones Cachés. En gras, le modèle d'ANN final choisi.

Modèle	CC	NC	**Decay**	**Sensibilité**	**Spécificité**	**% de bien classés**	AUC (ROC)
Effet neurones cachés							
	1	5	0	0,88	0,64	0,74	0,76
	1	10	0	0,75	0,64	0,68	0,69
	1	25	0	0,88	0,64	0,74	0,76
	1	35	0	0,88	0,64	0,74	0,76
	1	45	0	0,88	0,64	0,74	0,76
	1	50	0	0,88	0,64	0,74	0,76
Effet "decay" régularisation							
	1	25	0	0,88	0,64	0,74	0,76
	1	25	0,5	0,75	0,64	0,68	0,69
	1	**25**	**0,9**	**0,88**	**0,64**	**0,74**	**0,76**

- Comparaison des packages :

Différents packages proposent une application pour les ANN dans le logiciel R. Le package « neuralnet » et « nnet » sont les 2 plus simples et ont la même architecture : ce sont des réseaux de neurones non bouclés appelés « perceptrons multi-couches ». Ces 2 packages sont différents car ils ne proposent pas les mêmes paramètres : Le package « nnet » ne permet pas de choisir le nombre de couches : il est fixé à 1 alors que le package « neuralnet » permet de faire varier ce paramètre. En revanche le package « nnet » possède le paramètre « weight decay » qui représente le paramètre de régularisation pour éviter le sur-apprentissage.

Le package « nnet » dans « caret » permet aussi une sélection des meilleurs paramètres par apprentissage à l'aide de méthodes de rééchantillonnage incluses dans ce package (validation croisée ou bootstrapping). Tandis que pour le package « neuralnet » il faut choisir au préalable et fixé les paramètres du modèle.

Nous avons donc choisi le package « nnet » dans « caret » pour sa simplicité d'utilisation, sa bonne performance (cité ci-dessus), il inclut les techniques de ré-échantillonage et possède le paramètre de régularisation.

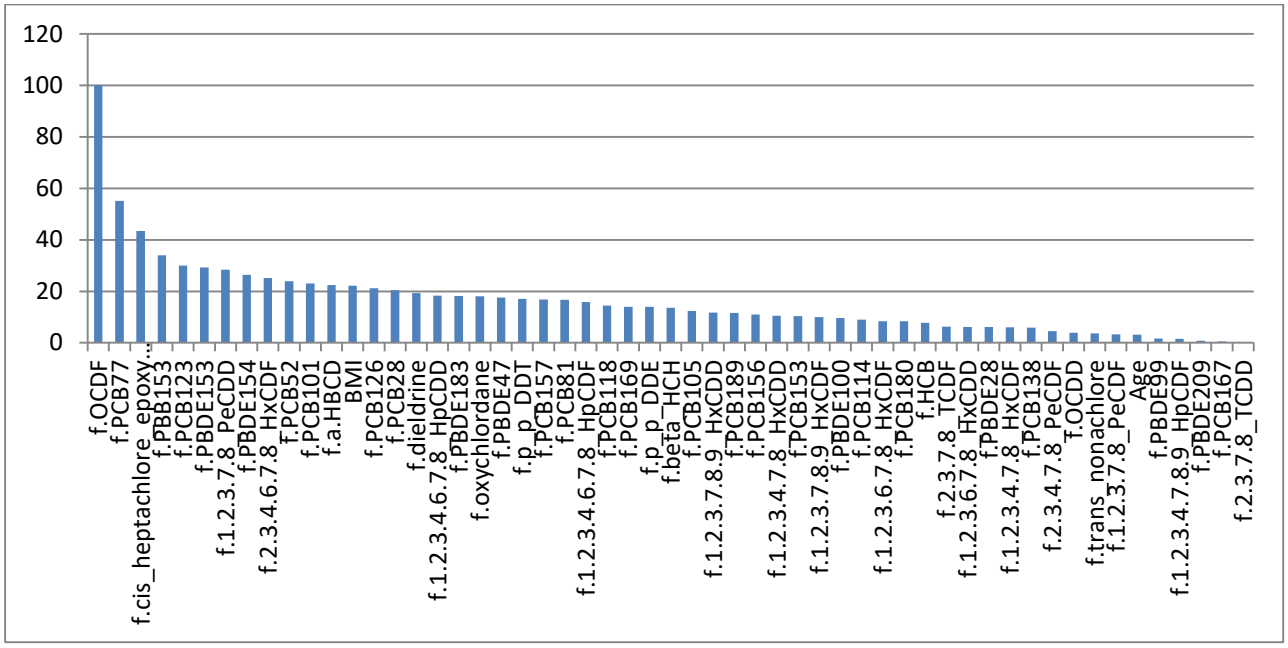
➢ Modèle final :

Aux vues de ces résultats, nous avons donc décidé de choisir comme modèle final celui à une couche cachée et 25 neurones cachés et 0.9 comme paramètre de régularisation. C'est une des solutions les plus performantes et ces paramètres sont très cohérents avec les estimations de la Formule 1 (Section III.3.a.ii). De plus, le paramètre de régularisation le plus élevé a été choisi pour éviter le risque de sur-apprentissage.

b) Importance des variables

Ici, dans la figure 7, est présentée l'importance des variables : le classement des variables par ordre d'importance d'association avec l'endométriose, échelonnées de 0 à 100. Les 5 premières variables les plus associées à l'endométriose selon les ANN sont dans l'ordre : l'OCDF, le PCB 77, le cis heptachlore époxyde, le PBB 153 et enfin le PCB 123. L'importance de ces variables a été extraite via les coefficients obtenus grâce à la formule d'Olden (Section III.3.a.iii).

Figure 7. Classement des variables par ANN en fonction de leur association avec l'endométriose



Nous avons également extrait les valeurs absolues des différents poids des variables d'exposition avec la formule d'Olden dans le Tableau A6 de l'Annexe : 21 polluants et l'IMC ont un coefficient inférieur à 1. Les 3 premières variables les plus associées à l'endométriose sont : l'OCDF avec un poids de 5,76, suivie par le PCB 77 avec un poids de 3.18, et enfin le cis heptachlore époxyde avec un poids de 2,50. Dans le Tableau A6 de l'annexe sont présentés les résultats des poids synaptiques de chaque variable de valeur absolue > 1.

2. Support Vector Machine

a) Optimisation et choix des paramètres

Lorsque l'on pratique une calibration automatique, en donnant une fourchette du paramètre C entre par exemple 1.e-4 à 1.e+4 et du paramètre sigma du noyau entre 1.e-4 et 1.e+4, le modèle choisi ayant les meilleures performances (AUC, spécificité, sensibilité) est celui qui a les valeurs les plus élevée de C (1.e+4) et les valeurs les plus faibles de sigma (1.e-4) : 68% d'individus bien classés (ayant une endométriose ou pas), une sensibilité à 74% et une spécificité de 88% (Tableau 4). En revanche, lorsque l'on fixe manuellement les paramètres, le meilleurs modèle est celui avec C = 1 et sigma = 0,01, avec 84 % d'individus bien classés et une aire sous la courbe ROC à 85%. Nous allons donc appliquer le meilleur modèle avec C = 1 et sigma = 0,01.

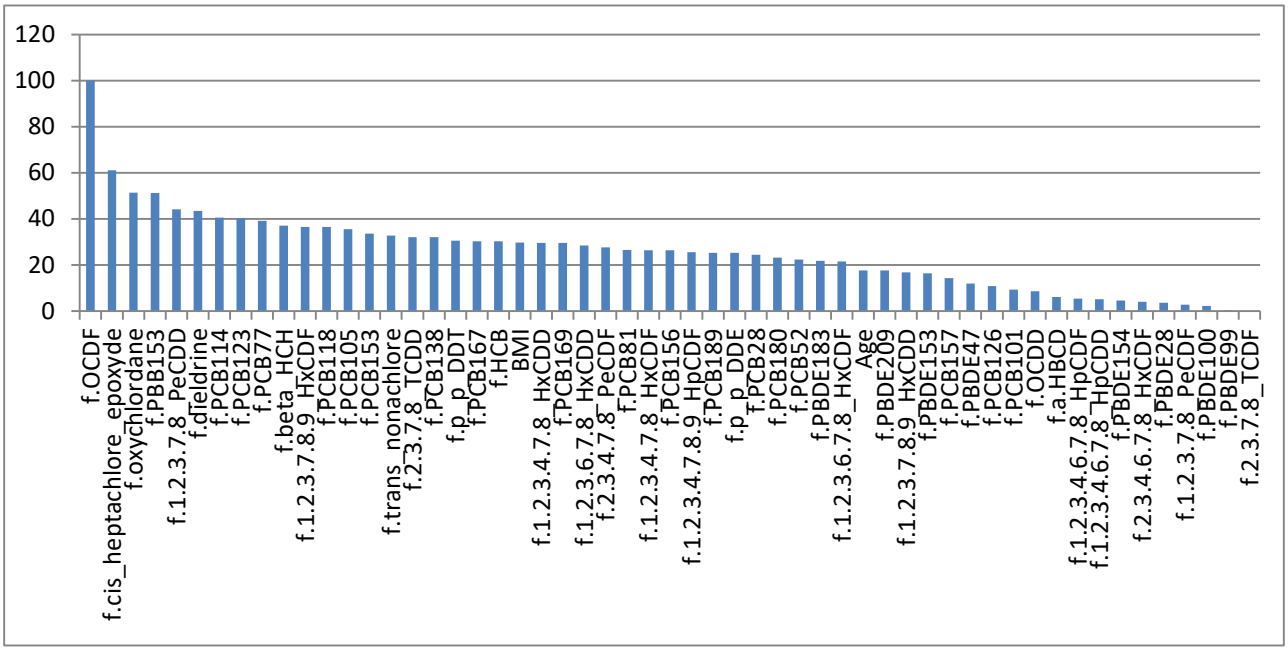
b) Importance des variables

Ce que l'on obtient dans les résultats une fois avoir appliqué l'algorithme choisi, ce sont des coefficients par vecteurs de supports qui sont difficilement interprétables. Pour rappel, les vecteurs de support sont les points se situant le plus proche de l'hyperplan établi. Nous pouvons cependant extraire les variables pertinentes, celles les plus associées au diagnostic de l'endométriose comme expliqué section III.3.b.iii. Parmi les 5 premières variables les plus importantes, on retrouve l'OCDF en premier comme pour les réseaux de neurones, le cis heptachlore époxyde en second, l'oxychlordane qui au contraire apparaît dans les variables non pertinentes des réseaux de neurones, le PBB 153 et le 1.2.3.7.8 PCDD. Par comparaison avec les réseaux de neurones parmi les 10 variables les plus importantes, 6 variables ressortent à la fois dans les SVM et dans les ANN et sont l'OCDF, le cis heptachlore époxyde, le PBB 153, le PCB 123, le PCB77, et enfin, la dioxine 1.2.3.7.8 PeCDD (Figure 8).

Tableau 4. Résultats de la performance de prédiction des SVM avec calibration des paramètres

Modèle	C	Sigma	Sensibilité	Spécificité	% de bien classés	AUC (ROC)
Automatique : avec C (de 1.e-4 à 1.e+4) et sigma = (1.e-4 à 1.e+4)						
	1.e+4	1.e-4	0,74	0,88	0,64	0,76
Automatique : avec C (de 1.e-5 à 1.e+5) et sigma = (1.e-5 à 1.e+5)						
	1.e+5	1.e-5	0,75	0,64	0,68	0,69
Manuellement						
Effet sigma						
	1	0,1	0,5	0,55	0,53	0,52
	1	**0,01**	**0,87**	**0,81**	**0,84**	**0,85**
	1	0,001	0,38	1	0,74	0,69
	1	0,0001	0,38	1	0,74	0,69
Effet C						
	10	0,01	0,75	0,73	0,74	0,74
	100	0,01	0,75	0,73	0,74	0,74
	1000	0,01	0,75	0,73	0,74	0,74
	0.1	0,01	0,75	0,81	0,79	0,79
	10	0,1	0,5	0,55	0,53	0,53
Effet sigma en fonction de C						
	10	**0,001**	**1**	**0,64**	**0,79**	**0,81**
	10	0,0001	0,38	1	0,74	0,69
	100	0,1	0,5	0,55	0,53	0,53
	100	0,001	0,87	0,64	0,74	0,76
	100	**0,0001**	**1**	**0,64**	**0,79**	**0,82**
	1000	0,1	0,5	0,55	0,53	0,53
	1000	0,001	0,87	0,64	0,74	0,76
	1000	0,0001	0,87	0,64	0,74	0,76

Figure 8. Classement des variables par les SVM en fonction de leur association avec l'endométriose



3. « Boosting Trees »

a) Optimisation et choix du paramètres

Lorsque l'on calibre les « arbres boostés », nous pouvons constater que la performance de prédiction sur le jeu de test est la même, peu importe le nombre d'itérations choisi. Ce modèle est par ailleurs très performant avec un AUC à 91%, un taux de bien classés à 89%, une sensibilité à 82% et une spécificité à 100%. Nous avons donc choisi le modèle avec 150 itérations, car il a été choisi comme le meilleur dans la recherche automatique.

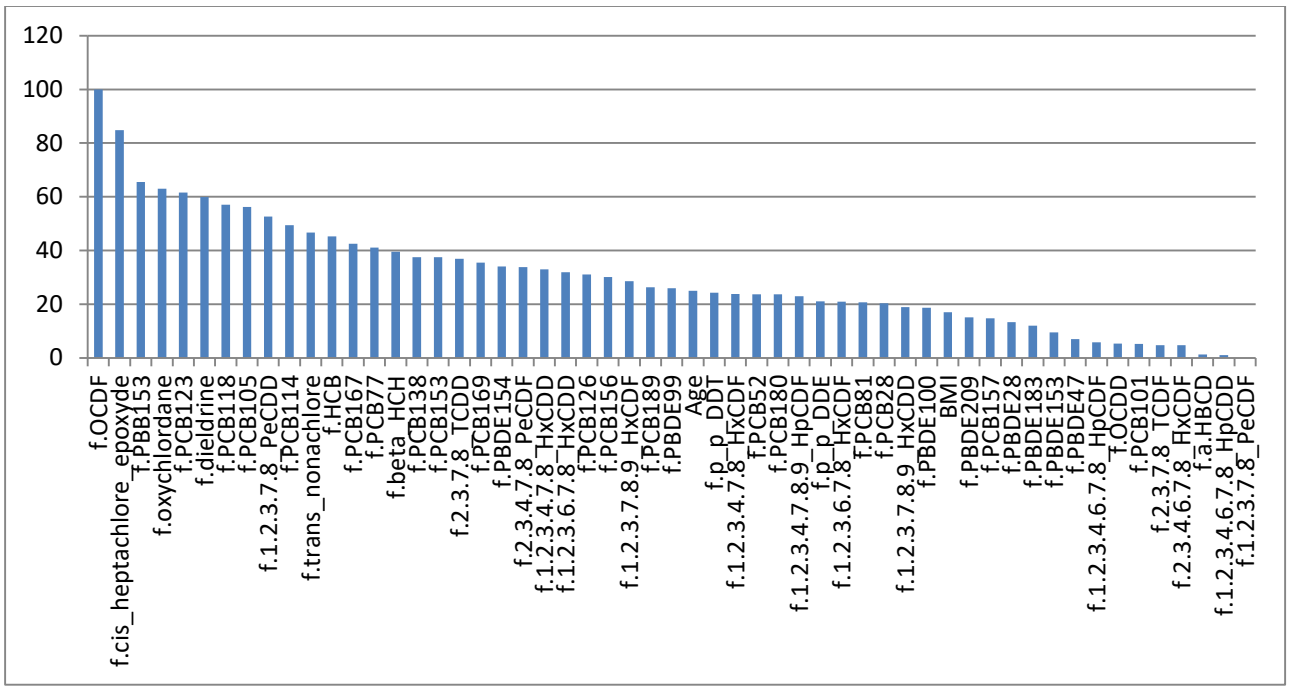
b) Importance des variables

Dans les résultats ici, les 5 premières variables les plus importantes, sont toujours l'OCDF en premier comme pour les ANN et les SVM, le cis heptachlore époxyde en second comme pour les SVM, s'en suit le PBB 153 retrouvé quant à lui en 4ème position pour les SVM et les ANN, puis l'oxychlordane qui apparaît aussi dans le top 5 des variables des SVM, mais qui au contraire apparaît dans les variables non pertinentes des ANN, puis enfin le PCB 123 retrouvé dans les 10 premières variables les plus associées à l'endométriose des 2 autres modèles (Figure 9). En revanche, le PCB 77 n'est pas retrouvé dans les arbres « boostés » comparé aux 2 autres précédents modèles. Au total, 5 polluants sont retrouvés parmi les 10 premières variables de chaque modèle : l'OCDF, le cis heptachlore époxyde, le PBB 153, le PBB 153, et enfin, la dioxine 1.2.3.7.8 PeCDD.

Tableau 5. Résultats de la performance de prédiction des « Boosting Trees » avec calibration des paramètres. Abréviations : nIter= nombre d'itérations

Modèle	nIter	Sensibilité	Spécificité	Taux de bien classés	AUC (ROC)
Automatique : Entre 50 et 150					
	150	0,82	1	0,89	0,91
Manuellement :					
	50	0,82	1	0,89	0,91
	100	0,82	1	0,89	0,91
	200	0,82	1	0,89	0,91
	300	0,82	1	0,89	0,91
	400	0,82	1	0,89	0,91
	500	0,82	1	0,89	0,91
	600	0,82	1	0,89	0,91
	1000	0,82	1	0,89	0,91

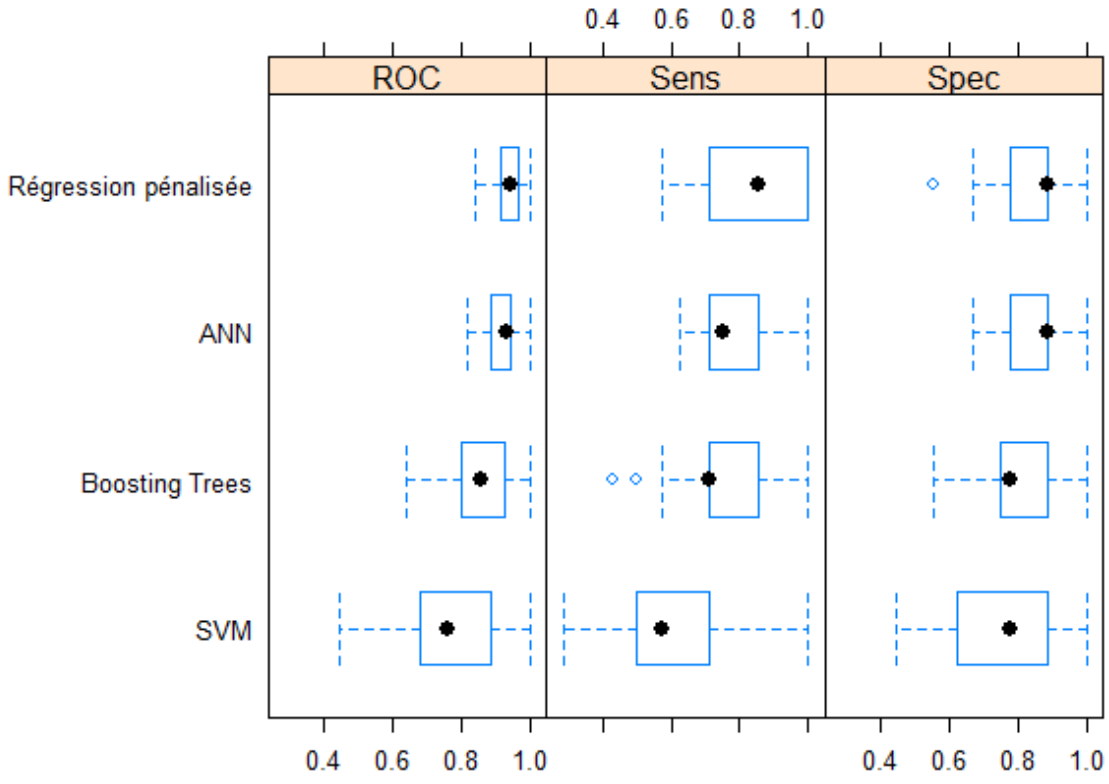
Figure 9. Classement des variables par les « Boosting Trees » en fonction de leur association avec l'endométriose



4. Comparaison des trois modèles

Nous avons comparé ces trois modèles entre eux en termes de performance de prédiction sur l'ensemble du jeu de données avec validation croisée (5 parties), et d'importance relative des variables. Nous les avons également comparés avec le modèle de régression logistique pénalisée type Elastic-net. Les résultats obtenus indiquent que c'est cette dernière méthode qui semble être la plus performante des quatre, suivi de près par le réseau de neurones avec 92% AUC, 79% de sensibilité et 86% de spécificité, et les « arbres boostés » avec 85% AUC, 76% de sensibilité et 82% de spécificité (Figure 10). En revanche, le SVM se situe un peu plus en dessous en termes de performance (77% AUC, 63% de sensibilité et 76% de spécificité), néanmoins, on peut voir que les étendues des distributions sont très larges pour ce modèle : 77% AUC avec une plage de [44%-89%], 63% de sensibilité avec une plage de [29%-71%] et enfin une spécificité à 76% avec une plage de [44%-89%]. Cela pourrait signifier que ce modèle est instable et l'on pourrait supposer qu'il y ait eu un risque de sur-apprentissage (« Over-fitting »). La moyenne des critères de performance des quatre modèles reste cependant élevée, ce qui signifie que ces modèles sont tous plutôt performants.

Figure 10. Comparaison de la performance de prédiction des trois méthodes d'apprentissage automatique sur l'ensemble du jeu de données avec la régression logistique pénalisée Elastic-Net (réalisée lors des analyses préliminaires). Abréviations : ANN = Réseau de neurones, SVM = Support vector machine, Sens = Sensibilité, Spec = spécificité.



Concernant l'importance relative des variables de chaque modèle, nous avons extrait les 10 premières variables les plus associées à l'endométriose (Tableau 6). Comme expliqué précédemment, l'importance de chaque variable est évaluée de manière différente, avec des approches et des formules différentes, d'une méthode à une autre. Il est donc difficile de pouvoir les comparer entre eux. C'est la raison pour laquelle nous allons évoquer ici seulement leur classement relatif. Les résultats des classements paraissent cohérents, et plusieurs variables ressortent dans chaque modèle notamment la concentration dans le tissu adipeux de l'OCDF, qui est la première variable la plus associée à l'endométriose dans tous les modèles, suivi par le cis-heptachlore-époxyde que l'on retrouve dans tous les modèles parmi les 4 premières variables. Le PCB 123 et la dioxine 1.2.3.7.8 PeCDD sont également retrouvés dans tous les modèles mais ne sont pas non plus classés de la même manière. Le PCB 77 se retrouve dans les premières variables pour trois modèles mais n'apparaît pas dans les 10 premières variables des « arbres boostés ». De même pour le PBB 153, que l'on voit dans tous les modèles sauf pour la régression logistique pénalisée elastic-net. On identifie donc un bloc de cinq polluants retrouvés dans les trois modèles appliqués comme les plus associés à l'endométriose : l'OCDF, le Cis heptachlore époxyde, le PBB 153, le PCB 123 et la dioxine 1.2.3.7.8 PeCDD.

Tableau 6. Comparaison du classement de l'importance des variables des 4 modèles

ANN	SVM	Boosting Trees	Elastic-Net régression
OCDF	OCDF	OCDF	OCDF
PCB 77	Cis heptachlore époxyde	Cis heptachlore époxyde	PCB 77
Cis heptachlore époxyde	oxychlordane	PBB 153	PBDE 153
PBB 153	PBB 153	oxychlordane	cis-heptachlore époxyde
PCB 123	1.2.3.7.8 PeCDD	PCB 123	PCB 81
PBDE 153	dieldrine	dieldrine	1.2.3.7.8 PeCDD
1.2.3.7.8 PeCDD	PCB 114	PCB 118	PCB123
PBDE154	PCB 123	PCB 105	PCB 101
2.3.4.6.7.8 HxCDF	PCB 77	1.2.3.7.8 PeCDD	2.3.4.6.7.8 HxCDF
PCB52	β HCH	PCB 114	PBDE 183

VIII. DISCUSSION

1. Polluants organiques persistants associés à l'endométriose et performance des modèles utilisés

Dans notre étude, nous avons utilisé trois méthodes alternatives d'apprentissage statistiques : les Réseaux de neurones artificiels (ANN), les « arbres boostés » (Boosting Trees) et les machine à vecteurs de supports (SVM), pour évaluer l'association entre un groupe large de polluants très corrélés et l'endométriose. Notre étude, basée sur des cas sévères d'endométriose, montrent que les niveaux internes de certains mélanges de polluants présents dans le tissu adipeux, seraient associés à l'endométriose, comme l'a démontré l'étude antérieure sur la même base de données avec un modèle statistique classique de régression logistique (9). La méthode la plus performante identifiée ici, en termes de capacité de prédiction est l'ANN, suivi des « Boosting Trees » puis en dernière, mais restant toujours performante, les SVM. Concernant l'importance des variables un bloc de cinq polluant sont identifiés comme les biomarqueurs d'exposition présents dans le tissu adipeux, les plus associés à l'endométriose avec les 3 modèles d'apprentissage statistiques étudiés. Il s'agit de la dioxine « Octachlorodibenzofurane » (OCDF), du pesticide organochloré « cis-heptachlore-époxyde », du « polychlorobiphényl (PCB) 123 », du retardateur de flamme bromé « polybromobiphényl (PBB) 153 » et de la dioxine « 1.2.3.7.8 Pentachlorodibenzoparadioxine (PeCDD) ». Comme l'OCDF se situe en première place pour chaque modèle, il peut être considéré comme le plus fortement associé à l'endométriose dans cette étude.

Ces résultats sont finalement cohérents avec l'étude antérieure réalisée sur la même base de données avec la méthode de régression logistique (9) : les associations significatives les plus fortes observées étaient l'OCDF avec un OR (IC95%) ajusté sur l'âge et le BMI de 4,60 (2,47 ; 9,84) suivi de près par PBB 153 : OR=4,10 (1,79 ; 10,89) et le cis-heptachlore-époxyde : 3,23 (1,81 ; 6,51). Le PCB 123 avait un OR de 1,99 (1,21 ; 3,47) et la dioxine 1.2.3.7.8 PeCDD était également associée de manière significative avec un OR de 1,79 (1,06 ; 3,21). Les résultats de cette étude sont également cohérents avec l'étude réalisées avec la régression logistique pénalisée Enet : parmi les polluants les plus associés étaient retrouvés l'OCDF et le Cis-heptachlore-epoxyde (29).

2. Forces de l’étude

a) Application des modèles dans ce cas d’étude

L’application de ces 3 méthodes en épidémiologie environnementale reste à ce jour limitée, ce qui constitue l’originalité de cette thèse. Ces méthodes apparaissent intéressantes car elles semblent être mieux adaptées à nos données que les modèles classiques de régression logistique, par le fait qu’elles puissent intégrer un grand nombre de variables et les évaluer toutes en même temps et non indépendamment des unes des autres, qu’elles s’affranchissent de la multicollinéarité des variables et enfin qu’elles puissent traiter de combinaisons non linéaires entre les variables. Concernant l’application de ces modèles par un épidémiologiste sans formation de bio-informaticien ou mathématicien, ni poussée en statistiques, une autre qualité de ces méthodes est leur facilité d’application sur le logiciel R grâce aux packages qui leur sont dédiés, notamment grâce au package « caret » qui facilite l’application des techniques de validation, validation croisée et rééchantillonnage. Un récapitulatif des forces et des faiblesses de chaque modèle figure dans le tableau 7.
La méthode alternative la plus facilement interprétable et la plus performante ici, est la régression logistique pénalisée Elastic-Net utilisée dans les études préliminaires, car les coefficients obtenus s’interprètent comme l’odds ratio. Il serait donc pertinent d’utiliser cette méthode en épidémiologie environnementale en attendant de mieux connaître les autres modèles complexes. Cependant, cette méthode ne traite que de combinaisons linéaires entre les variables, et il existe un risque de biais plus ou moins important en fonction de la valeur prise pour le paramètre de pondération (« shrinkage »). Certaines études conseillent même d’appliquer ces méthodes comme une première étape de sélection des variables avant d’appliquer les modèles plus usuels ce qui pourrait être une bonne perspective d’utilisation (30). Ces différents modèles sont donc très complémentaires en termes d'interprétabilité et de performance pour identifier les associations linéaires et non linéaires. Ainsi, la combinaison d'une batterie d'algorithmes peut être une approche efficace pour explorer des ensembles de données complexes sans connaître la structure sous-jacente. Par exemple, dans le cas d'associations non linéaires, SVM ou ANN surclasseraient les algorithmes linéaires (ENET).
Ce présent travail a donc permis d'établir une base méthodologique afin d’intégrer des données multidimensionnelles à partir de plateformes métabolomiques et « exposomiques », pour mieux comprendre les associations fonctionnelles entre les expositions et l'endométriose (c'est-à-dire le projet EndoxOmic).

b) Applications de ces modèles à d'autres études

Malgré l'utilisation émergente de modèles multi polluants en épidémiologie environnementale, peu d'études ont appliqué des algorithmes d'apprentissage statistique pour mieux comprendre les associations complexes exposome-santé (44). Concernant l'endométriose, une étude (68) de 2012 a appliqué une approche de réduction de dimension des données : le réseau Bayésien, pour identifier les biomarqueurs en tenant compte des covariables biologiquement pertinentes, les plus associés à l'endométriose.
Parallèlement, plusieurs études récemment publiées ont utilisé une série de modèles d'apprentissage automatique pour étudier d'autres facteurs de risque liés à une problématique de santé. Zhao et al., 2019 (69) ont testé quatre algorithmes différents (ANN, SVM, Adaboost, Random Forest (RF)) sur une population de 1113 travailleurs exposés au bruit industriel afin de prédire les déficiences auditives. La précision prédictive se situait entre 78,6 et 80,1% pour les quatre modèles, ce qui est comparable aux précisions prédictives trouvées dans cette étude. Une autre étude de cohorte menée par les équipes de cancérologie digestive des hôpitaux de Paris (40), avec 49 hommes et 46 femmes, utilise des réseaux de neurones profonds et les SVM afin de prédire la réponse thérapeutique à une radio-chimiothérapie préopératoire chez des patients suivis pour un cancer du rectum : Le réseau de neurone prédit une réponse complète avec une précision de 80%, meilleure que celle du modèle de régression linéaire (69,5%) et du modèle SVM (71,58%). Ces résultats peuvent aider à identifier les patients qui pourraient bénéficier d'un traitement conservateur plutôt que d'une résection radicale.

3. Limites de l'étude

a) Liées aux données

Cependant, beaucoup de limites sont à émettre concernant notre étude. La première concerne la taille de l'échantillon : ici, le nombre d'observations de l'ensemble de données est de 99, une taille qui sans être très faible peut ne pas être suffisante pour assurer la stabilité des modèles. Or ces modèles sont généralement utilisés sur des échantillons de très grande taille. La taille de l'échantillon peut également avoir une incidence sur la stabilité des coefficients et la reproductibilité des résultats, ce qui est un problème inhérent aux étalonnages basés sur les données et sur le facteur de variation pour sélectionner les paramètres de réglage (70). Il n'existe pas aujourd'hui de consensus dans la littérature sur le nombre minimum d'observations requis pour exécuter ces modèles, cela dépend de la base de données ainsi que des modèles utilisés. Par ailleurs, il existe dans la littérature, des études utilisant ces modèles et ayant des petits effectifs, avec une performance non altérée. Des travaux complémentaires devraient permettre d'évaluer l'impact du nombre d'observations sur la performance des modèles, comme des études de simulation par exemple. De plus, dans notre étude de cas, pouvoir disposer d'une

grande base de données externe sur l'endométriose aurait pu faire office de validation externe des algorithmes, or ici nous n'avions qu'un seul jeu de données que nous avons séparé pour entraîner l'algorithme.

b) Liées aux modèles

Par ailleurs, la flexibilité de ces modèles, c'est-à-dire leur capacité à s'adapter à chaque donnée individuelle, peut entraîner un risque de se sur-ajuster aux données étudiées et donc un plus haut risque d'instabilité et une difficulté de généraliser les résultats à la population générale. Il est donc nécessaire de trouver un compromis avec cette flexibilité et l'instabilité du modèle (71) (45) (30). Une autre partie importante que nous n'avons pas abordé dans cette étude est l'interaction entre variables, car ces modèles sont aussi très prometteurs dans le fait qu'ils puissent trouver les interactions et les intégrer dans le modèle (72) (45).

L'application plus poussée et consolidée sur le plan du niveau de confiance de ces modèles demande toutefois un renforcement préalable des connaissances théoriques afférentes en statistique, mathématique et bioinformatique. Une difficulté supplémentaire est l'existence d'une multitude d'algorithmes et de fonction différentes pour chaque méthode : par exemple, le réseau de neurones à plusieurs couches et fonctions d'activation comme le perceptron multicouches, les réseaux bouclés, les neurones sommateurs, polynomiaux, … parmi lesquels un choix doit être opéré pour aboutir à un paramétrage le plus adapté possible à la question de recherche traitée et aux données support. Des connaissances approfondies sont également nécessaires pour la calibration des paramètres des modèles qui est une étape primordiale dans l'application de ces modèles. Cette étape reste complexe et assez chronophage et peut avoir un gros impact sur les résultats finaux, en termes de prédiction comme d'évaluation de l'importance des variables. Elle est cruciale pour éviter les risques de sur-apprentissage qui fait partie d'une des principales limites des approches de l'apprentissage automatique.

En outre, cette étude n'a étudié que quelques modèles d'apprentissage automatique supervisés, et il existe beaucoup d'autres méthodes visant à s'affranchir des problématiques liées à ces données, notamment les modèles Bayésien non abordés ici, des méthodes d'équations structurelles de données mais aussi les méthodes d'apprentissage automatique non supervisées, qui auraient pu être également une bonne alternative aux méthodes conventionnelles.

c) Liées à l'interprétabilité

Du point de vue épidémiologique, l'interprétation des résultats de ces méthodes peut paraître complexe, car elles ne produisent pas de coefficients similaires à l'Odds Ratio permettant de mesurer la force de l'association, ni de p-value et d'intervalles de confiance pour mesurer la fiabilité du modèle. On obtient des coefficients échelonnés entre 0 et 100, mais nous ne pouvons les interpréter comme un odds ratio. C'est la raison pour laquelle ces méthodes restent en l'état encore difficilement interprétables, leurs résultats s'éloignant du langage de l'épidémiologie classique. De plus, les variables sont ordonnées à l'aide de formules et démarches différentes pour les trois approches considérées, d'autant plus qu'il peut exister plusieurs formules pour une même approche, comme discuté précédemment dans le cas des ANN. Il est donc difficile de savoir quelle formule appliquer pour classer les variables, et cela nécessite une connaissance théorique des démarches sous-jacentes. Dans le cas des ANN, on peut trouver dans la littérature, des exemples de traduction de l'algorithme pour pouvoir le comparer avec la régression logistique afin de réduire l'effet « boîte noire » de ce modèle. Dans l'article « Epidemiologic interpretation of artificial neural networks », les auteurs pensent que les ANN peuvent être très utile en épidémiologie, notamment pour des résoudre des problèmes complexes de classification (73). Il a été également montré que les algorithme de « boosting » pour les problèmes de classification binaires ont un lien avec le modèle de régression logistique (74). Afin de surmonter ces limites, le présent travail a été amélioré en couplant un algorithme destiné à prendre en charge la sélection variable de classificateurs, afin d'obtenir la sélection de l'ensemble minimal variables explicatives pour minimiser le taux d'erreur de classification. Dans tous les cas, ces approches doivent être envisagées pour soutenir la réduction de dimension et la sélection de variables dans un contexte de dimension élevée, préalable à une inférence statistique (recherche des facteurs de risques).

4. Perspectives

a) Epidémiologie environnementale

La possibilité d'effets à faibles doses de certains agents environnementaux résultant d'expositions précoces, chez le jeune enfant mais aussi *in utero*, voire avant la conception par altération génétique ou épigénétiques des gamètes, est aujourd'hui un sujet central de préoccupation et de recherche en santé environnementale. L'épidémiologie se débat de plus en plus avec des problèmes d'expositions corrélées et de risques relatifs faibles. En conséquence, certains chercheurs ont fortement mis l'accent sur l'épidémiologie moléculaire, alors que d'autres ont plaidé pour l'importance du contexte de la population et de la réintégration de l'épidémiologie dans la santé publique. L'épidémiologie environnementale présente plusieurs caractéristiques uniques qui rendent ces débats particulièrement

pertinents. Le très grand nombre d'expositions environnementales nécessite une hiérarchisation des priorités et les risques relatifs sont généralement très faibles. En outre, de nombreuses expositions environnementales ne peuvent être abordées qu'en comparant les populations plutôt que les individus, et la perturbation des écosystèmes locaux et globaux nous oblige à développer de nouvelles méthodes de conception d'étude (75). Toutes ces questions, et d'autres non citées ici, n'auraient pas pu être soulevées à partir des seules méthodes épidémiologiques classiques. Ces modèles d'apprentissage automatique constituent donc une grande opportunité en épidémiologie environnementale. Nous pouvons, d'une part, mesurer les capacités de prédiction de ces modèles avec des indicateurs spécifiques et, d'autre part, extraire un classement des variables les plus associées à la problématique de santé pour chaque modèle. Néanmoins, les chercheurs peuvent être réticents à utiliser ces méthodes à cause de la difficulté d'interprétation des résultats, notamment pour faire des inférences statistiques directes, et du fait de laisser l'ordinateur choisir le meilleur modèle (76). C'est pourquoi, il apparaît évident qu'il faudra, pour progresser en santé environnementale, aussi bien du point de vue des connaissances que de la prévention, mettre en place des structures qui favorisent la confrontation des approches (épidémiologiques, toxicologiques, médicales, biologiques, génétiques, métrologiques, sociologiques, etc.) et la multidisciplinarité, comme cela est fait dans d'autres pays. La nécessité de cette évolution a été bien identifiée au cours de l'élaboration du deuxième Plan national santé environnement et, plus encore peut-être, dans certaines réflexions stratégiques récentes (77). Le défi sera de savoir comment exploiter au mieux ces technologies pour des objectifs de santé environnementale. Avec la disponibilité croissante des données, le besoin de plans d'étude robustes et nouveaux pour identifier les associations de causalité. La distinction entre population et l'évaluation individuelle de l'exposition et des résultats s'érodera en raison de la magnitude des données individuelles disponibles via des capteurs interfacés avec les technologies intelligentes et les biotechnologies. L'épidémiologiste environnemental de demain aura besoin de compétences différentes et perfectionnées pour travailler efficacement dans toutes les disciplines, poser les bonnes questions, mettre en œuvre méthodes de conception et d'analyse les plus appropriées pour identifier les causes relations, susciter et engager l'intérêt du public et de la politique décideurs et concevoir des interventions efficace (78).

b) Ouverture à la santé au sens général

Ces changements pourraient avec grand profit diffuser en santé publique « classique » et dans l'enseignement de celle-ci auprès de l'ensemble des professionnels de santé, en particulier les médecins. Le recentrage, tant souhaité par les pionniers de la santé publique, de la médecine de soins sur la médecine de prévention passera donc peut-être par le développement bien compris de la santé environnementale, au bénéfice de tous les acteurs et de la population (77).

Au-delà de l'utilisation de ces algorithmes en épidémiologie, l'intelligence artificielle (IA) ouvre la voie d'une médecine où le suivi en temps réel du patient et des traces qu'il produit (suivi de son état

physiologique, description de ses symptômes, interactions avec son environnement…) est essentiel pour entraîner et améliorer en continu la fiabilité des techniques d'IA utilisées à des fins médicales.
Le développement de l'IA est amené à transformer en profondeur les pratiques des professionnels de santé : aide au diagnostic, appui à la construction d'une thérapie, suivi évolutif du patient… S'il n'est pas question de remplacer les médecins par la machine, l'enjeu est bien d'organiser des interactions vertueuses entre l'expertise humaine et les apports de l'IA dans l'exercice quotidien de la médecine. L'intégration de l'IA pour analyser les données massives des plates-formes à haut débit pour le profilage moléculaire (métabolomique, protéomique) des patients permet un niveau sans précédent de phénotypage personnalisé et de développement de stratégies préventives. Par exemple, l'application de ces algorithmes avec le profil métabolomique de l'endométriose peut aider à identifier les biomarqueurs précliniques précoces raccourcissant la grande fenêtre actuelle entre les premiers symptômes et la confirmation chirurgicale (environ 7 à 10 ans).
L'appropriation des technologies médicales basées sur l'IA dans les pratiques de médecine va entraîner une réorganisation des professions médicales. Dans son rapport, Cédric Villani émet des recommandations concernant l'évolution des pratiques médicales : transformer les voies d'accès aux études de médecine : d'une part pour diversifier les profils et intégrer davantage d'étudiants spécialisés dans le domaine de l'informatique et de l'IA (création de double cursus, reconnaissance d'équivalence), et d'autre part pour mettre un terme à la logique de compétition tout au long du cursus universitaire qui s'avère contre-productive pour développer une coordination transdisciplinaire et structure les postures d'autorité médecins - patients ; former les professionnels de santé aux usages de l'intelligence artificielle, de l'IOT et du big data en santé, ainsi qu'aux compétences de coordination, d'empathie et du rapport avec les patients (ex. : expérience virtuelle pour mieux voir, comprendre la vie des patients). Cette transformation de la formation initiale pourrait avoir lieu dans la réforme en cours du premier et deuxième cycle de médecine, etc. (34).

Tableau 7. Récapitulatif des forces et des faiblesses de chaque modèle.

Modèle	Forces	Faiblesses
ANN	S'utilise facilement avec les packages R « nnet », et par l'intermédiaire du package « caret ». Permet de traiter les combinaisons non linéaires.	Difficultés d'interprétation. Plusieurs paramètres à calibrer. Beaucoup d'algorithmes différents disponibles. Nécessite des connaissances statistiques et bioinformatiques. Il existe plusieurs formules différentes pour extraire l'importance des variables du modèle.
SVM	Peut éviter le sur-apprentissage. Permet de traiter les combinaisons non linéaires. S'utilise facilement avec la fonction « svm » dans le package « e1071 » et par l'intermédiaire du package « caret » dans R.	Difficultés d'interprétation Plusieurs paramètres à calibrer Pas de formule spécifique pour extraire l'importance des variables, seulement sur le critère de performance AUC Nécessite des connaissances statistiques et bioinformatique
Boosting Trees	Un seul paramètre à calibrer avec l'algorithme « AdaBoost-M1 » S'utilise facilement avec le package « fastAdaboost » et par l'intermédiaire du package « caret » dans R. Permet de traiter les combinaisons non linéaires. Algorithme « AdaBoost-M1 » appuyé dans la littérature	Difficultés d'interprétation Beaucoup d'algorithmes différents disponibles et récents plus performant que celui-ci par ailleurs. Nécessite des connaissances statistiques et bioinformatique

IX. CONCLUSION

Il est à présent reconnu que le facteur environnemental contribue à l'étiologie d'un grand nombre de pathologies. Alors que les études de surveillance montrent que nous sommes exposés à une multitude de produits chimiques potentiellement impliqués dans ces effets néfastes, mais la caractérisation de l'impact sur la santé de ces polluants en mélange complexe reste un défi majeur. En effet, d'une part les modèles statistiques classiques de l'épidémiologie environnementale ne permettent pas d'évaluer ce « cocktail » de polluants et d'autre part, la forte colinéarité généralement observée entre ces mesures de polluants constitue un frein dans l'utilisation de nombreuses méthodes statistiques qui sont sensibles à la multicolinéarité. Au regard de la littérature, il existe maintenant un nombre croissant de méthodes alternatives d'apprentissage automatique accessibles, avec chacune leurs avantages et leurs inconvénients, pour pouvoir remédier à cette problématique et étudier l'association entre ces mélanges de polluants et une pathologie définie. Ainsi, il est maintenant prouvé qu'il existe une relation entre certains polluants, perturbateurs endocriniens et les pathologies gynécologiques (Gore et al., 2015). Dans le cadre de notre étude, nous nous sommes intéressés à la mise en évidence de certains polluants en relation avec l'endométriose. Cette dernière est une pathologie gynécologique fréquente atteignant 5 à 10% des femmes en âge de procréer dont 35 à 50% présentent des douleurs pelviennes associées voire une infertilité (Missmer et al., 2004). L'étiologie environnementale de cette pathologie a déjà été démontré depuis quelques années par des modèles statistiques classiques ou linéaires. Dans notre étude, nous avons étudié l'association entre les niveaux internes de polluants présents dans le tissu adipeux et la présence d'endométriose. L'originalité de notre étude réside dans la recherche de modèles « multi-polluants » construits sur la base de plusieurs méthodes d'apprentissage automatique qui sont : les réseaux de neurones, les SVM, et les arbres boostés. La méthode la plus performante identifiée ici, en termes de capacité de prédiction sont les ANN, suivi des « Boosting Trees » puis en dernière, mais restant toujours performante, les SVM. Celles-ci ont également été comparée à la capacité prédictive d'un modèle plus usuel qu'est la régression logistique pénalisée ElasticNet, qui se retrouve être la plus performante. Les trois approches appliquées ont mis en évidence cinq polluants en commun les plus associés à l'endométriose : l'OCDF, le pesticide organochloré cis-heptacholre-époxyde, le PCB 123, le PBB 153 et le 1.2.3.7.8 PeCDD. Ces derniers sont également ressortis comme fortement associés à l'endométriose avec un modèle de régression logistique simple et pénalisé, par l'approche Elastic-Net dans les études préliminaires. Ces approches d'apprentissage automatique représentent donc un enjeu majeur en épidémiologie environnementale pour permettre de prendre en compte ces mélanges de polluants et de se rapprocher de scénarios d'exposition plus réalistes. Cependant, elles sont encore mal comprises et cela nécessite un approfondissement des connaissances biostatistiques et bioinformatiques de ces modèles qui demandent à faire l'objet d'études ultérieures. Ces avancées méthodologiques pourraient faire partie intégrante de l'enseignement en épidémiologie et nécessitent un travail pluridisciplinaire avec les biostatistques et la bioinformatique. Au-delà de l'application de ces méthodes à l'épidémiologie, inscrire l'intelligence artificielle dans les cursus

théorique et pratique des métiers de la santé pourraient être une bonne suggestion afin de faire avancer les progrès de la médecine et d'améliorer les traitements des données de santé.

X. BIBLIOGRAPHIE

1. Wild CP. The exposome: from concept to utility. Int J Epidemiol. févr 2012;41(1):24-32.

2. Ministère des solidarités et de la santé. Plan national Santé-Environnement 4 (PNSE 4), « Mon environnement, ma santé » (2020-2024) [Internet]. 2020. Disponible sur: https://solidarites-sante.gouv.fr/sante-et-environnement/les-plans-nationaux-sante-environnement/article/plan-national-sante-environnement-4-pnse-4-mon-environnement-ma-sante-2020-2024

3. Commissariat général au développement durable, Santé Publique France. Focus Environnement & santé. In: L'environnement en France [Internet]. La documentation française. 2019. Disponible sur: https://www.statistiques.developpement-durable.gouv.fr/sites/default/files/2019-07/ree2019-focus-environnement-sante-juillet2019_0.pdf

4. Fontanet A. Histoire de l'épidémiologie [Internet]. Chaire santé publique - Collège de France; 2018. Disponible sur: https://www.college-de-france.fr/site/arnaud-fontanet/p4605062636254075_content.htm

5. Santé Publique France. Cohorte Elfe [Internet]. En cours. Disponible sur: https://www.santepubliquefrance.fr/etudes-et-enquetes/cohorte-elfe

6. Santé Publique France. Esteban [Internet]. Disponible sur: https://www.santepubliquefrance.fr/etudes-et-enquetes/esteban

7. Enquête AGRICAN : les agriculteurs en meilleure santé que le reste de la population. Pour. 2012;214(2):73.

8. Ploteau S. Etude du lien entre l'exposition aux polluants organiques persistants et l'endométriose [Internet]. 2016. Disponible sur: http://www.theses.fr/2016ONIR087F/document

9. Ploteau S, Cano-Sancho G, Volteau C, Legrand A, Vénisseau A, Vacher V, et al. Associations between internal exposure levels of persistent organic pollutants in adipose tissue and deep infiltrating endometriosis with or without concurrent ovarian endometrioma. Environ Int. 2017;108:195-203.

10. Sampson JA. Peritoneal endometriosis due to the menstrual dissemination of endometrial tissue into the peritoneal cavity. Am J Obstet Gynecol. 1927;14(4):422-69.

11. Inserm. Endométriose [Internet]. Disponible sur: https://www.inserm.fr/information-en-sante/dossiers-information/endometriose

12. Giudice LC. Clinical practice. Endometriosis. N Engl J Med. 24 juin 2010;362(25):2389-98.

13. Kopelman A, Girão MJBC, Bonetti TCS, Carvalho CV, da Silva IDCG, Schor E. Analysis of Gene Expression in the Endocervical Epithelium of Women With Deep Endometriosis. Reprod Sci Thousand Oaks Calif. 2016;23(9):1269-74.

14. Dun EC, Taylor RN, Wieser F. Advances in the genetics of endometriosis. Genome Med. 14 oct 2010;2(10):75.

15. Gore AC, Chappell VA, Fenton SE, Flaws JA, Nadal A, Prins GS, et al. Executive Summary to EDC-2: The Endocrine Society's Second Scientific Statement on Endocrine-Disrupting Chemicals. Endocr Rev. déc 2015;36(6):593-602.

16. Smarr MM, Kannan K, Buck Louis GM. Endocrine disrupting chemicals and endometriosis. Fertil Steril. sept 2016;106(4):959-66.

17. Bruner-Tran KL, Osteen KG. Dioxin-like PCBs and endometriosis. Syst Biol Reprod Med. avr 2010;56(2):132-46.

18. Heilier J-F, Donnez J, Lison D. Organochlorines and endometriosis: A mini-review. Chemosphere. mars 2008;71(2):203-10.

19. Buck Louis GM. Early Origins of Endometriosis: Role of Endocrine Disrupting Chemicals. In: Giudice LC, Evers JLH, Healy DL, éditeurs. Endometriosis [Internet]. Oxford, UK: Wiley-Blackwell; 2012 [cité 25 nov 2019]. p. 153-63. Disponible sur: http://doi.wiley.com/10.1002/9781444398519.ch15

20. Cano-Sancho G, Ploteau S, Matta K, Adoamnei E, Louis GB, Mendiola J, et al. Human epidemiological evidence about the associations between exposure to organochlorine chemicals and endometriosis: Systematic review and meta-analysis. Environ Int. 2019;123:209-23.

21. INERIS. Données technico-économique sur les substances chimiques en France : les Dioxines. 2006.

22. Fiedler H. [Origin, structure and distribution of dioxins]. DTW Dtsch Tierarztl Wochenschr. août 2006;113(8):304-7.

23. Cancer-environnement. Dioxines et risques de cancer [Internet]. Disponible sur: https://www.cancer-environnement.fr/367-Dioxines.ce.aspx

24. INERIS. Données technico-économique sur les substances chimiques en France : les PolyChloroBiphényls (PCBs) [Internet]. 2011 p. 89. Disponible sur: https://rsde.ineris.fr/

25. European Food Safety Authority (EFSA). Retardateurs de flamme bromés [Internet]. Disponible sur: https://www.efsa.europa.eu/fr/topics/topic/brominated-flame-retardants

26. Porta M, Zumeta E. Implementing the Stockholm Treaty on Persistent Organic Pollutants. Occup Environ Med. oct 2002;59(10):651-2.

27. Aligon D, Bonneau J, Garcia J, Gomez D, Le Goff D. Mémoire - Estimation des expositions de la population générale aux insecticides : les Organochlorés, les Organophosphorés et les Pyréthrinoïdes. Ecole des Hautes Etudes en Santé Publique (EHESP); 2010.

28. Ploteau S, Antignac J-P, Volteau C, Marchand P, Vénisseau A, Vacher V, et al. Distribution of persistent organic pollutants in serum, omental, and parietal adipose tissue of French women with deep infiltrating endometriosis and circulating versus stored ratio as new marker of exposure. Environ Int. déc 2016;97:125-36.

29. Cano-Sancho G, Ploteau S, Volteau C, Legrand A, Vénisseau A, Vacher V, et al. the associations between persistent organic pollutants and endometriosis: single versus multi-pollutant approaches applied to a first case-control study. In 2017.

30. Sun Z, Tao Y, Li S, Ferguson KK, Meeker JD, Park SK, et al. Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons. Environ Health. déc 2013;12(1):85.

31. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning [Internet]. New York, NY: Springer New York; 2009 [cité 25 nov 2019]. (Springer Series in Statistics). Disponible sur: http://link.springer.com/10.1007/978-0-387-84858-7

32. Kouwaye B. Contributions of statistical learning to GLMM and LASSO methods: Application to statistical modeling of malaria morbidity at Tori-Bossito (Benin) [Internet]. UAC - University of Abomey Calavi; 2018. Disponible sur: https://hal.archives-ouvertes.fr/tel-01736933

33. Mitchell TM. Machine Learning. New York: McGraw-Hill; 1997. 414 p. (McGraw-Hill series in computer science).

34. Villani C, Schoenauer M, Bonnet Y, Berthet C, Cornut A-C. Donner un sens à l'intelligence artificielle : Pour une stratégie nationale et européenne. [Internet]. 2018 p. 195-203. Disponible sur: https://www.aiforhumanity.fr/

35. Liu S, Liu S, Cai W, Pujol S, Kikinis R, Feng D. Early diagnosis of Alzheimer's disease with deep learning. In: 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI) [Internet]. Beijing, China: IEEE; 2014 [cité 21 nov 2019]. p. 1015-8. Disponible sur: http://ieeexplore.ieee.org/document/6868045/

36. Cheng J-Z, Ni D, Chou Y-H, Qin J, Tiu C-M, Chang Y-C, et al. Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans. Sci Rep. avr 2016;6(1):24454.

37. Dilsizian SE, Siegel EL. Artificial Intelligence in Medicine and Cardiac Imaging: Harnessing Big Data and Advanced Computing to Provide Personalized Medical Diagnosis and Treatment. Curr Cardiol Rep. janv 2014;16(1):441.

38. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA. 13 déc 2016;316(22):2402.

39. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. févr 2017;542(7639):115-8.

40. Bibault J-E, Giraud P, Housset M, Durdux C, Taieb J, Berger A, et al. Deep Learning and Radiomics predict complete response after neo-adjuvant chemoradiation for locally advanced rectal cancer. Sci Rep. déc 2018;8(1):12611.

41. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. déc 2017;2(4):230-43.

42. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer. 2018;18(8):500-10.

43. Taylor KW, Joubert BR, Braun JM, Dilworth C, Gennings C, Hauser R, et al. Statistical Approaches for Assessing Health Effects of Environmental Chemical Mixtures in Epidemiology: Lessons from an Innovative Workshop. Environ Health Perspect [Internet]. déc 2016 [cité 25 nov 2019];124(12). Disponible sur: https://ehp.niehs.nih.gov/doi/10.1289/EHP547

44. Stafoggia M, Breitner S, Hampel R, Basagaña X. Statistical Approaches to Address Multi-Pollutant Mixtures and Multiple Exposures: the State of the Science. Curr Environ Health Rep. déc 2017;4(4):481-90.

45. 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.

46. Vriens A, Nawrot TS, Baeyens W, Den Hond E, Bruckers L, Covaci A, et al. Neonatal exposure to environmental pollutants and placental mitochondrial DNA content: A multi-pollutant approach. Environ Int. sept 2017;106:60-8.

47. Lenters V, Portengen L, Rignell-Hydbom A, Jönsson BAG, Lindh CH, Piersma AH, et al. Prenatal Phthalate, Perfluoroalkyl Acid, and Organochlorine Exposures and Term Birth Weight in Three Birth Cohorts: Multi-Pollutant Models Based on Elastic Net Regression. Environ Health Perspect. mars 2016;124(3):365-72.

48. Beckerman BS, Jerrett M, Martin RV, van Donkelaar A, Ross Z, Burnett RT. Application of the deletion/substitution/addition algorithm to selecting land use regression models for interpolating air pollution measurements in California. Atmos Environ. oct 2013;77:172-7.

49. Agay-Shay K, Martinez D, Valvi D, Garcia-Esteban R, Basagaña X, Robinson O, et al. Exposure to Endocrine-Disrupting Chemicals during Pregnancy and Weight at 7 Years of Age: A Multi-pollutant Approach. Environ Health Perspect. oct 2015;123(10):1030-7.

50. Berg V, Nøst TH, Pettersen RD, Hansen S, Veyhe A-S, Jorde R, et al. Persistent Organic Pollutants and the Association with Maternal and Infant Thyroid Homeostasis: A Multipollutant Assessment. Environ Health Perspect. janv 2017;125(1):127-33.

51. Gass K, Klein M, Sarnat SE, Winquist A, Darrow LA, Flanders WD, et al. Associations between ambient air pollutant mixtures and pediatric asthma emergency department visits in three cities: a classification and regression tree approach. Environ Health. déc 2015;14(1):58.

52. Sherriff A, Ott J, the ALSPAC Study Teama. Artificial neural networks as statistical tools in epidemiological studies: analysis of risk factors for early infant wheeze. Paediatr Perinat Epidemiol. nov 2004;18(6):456-63.

53. Zheng W, Tian D, Wang X, Tian W, Zhang H, Jiang S, et al. Support vector machine: Classifying and predicting mutagenicity of complex mixtures based on pollution profiles. Toxicology. nov 2013;313(2-3):151-9.

54. Cybenko G. Approximation by superpositions of a sigmoidal function. Math Control Signals Syst. déc 1989;2(4):303-14.

55. Guang-Bin Huang. Learning capability and storage capacity of two-hidden-layer feedforward networks. IEEE Trans Neural Netw. mars 2003;14(2):274-81.

56. Sheela KG, Deepa SN. Review on Methods to Fix Number of Hidden Neurons in Neural Networks. Math Probl Eng. 2013;2013:1-11.

57. Venables WN, Ripley BD. Modern Applied Statistics with S [Internet]. New York, NY: Springer New York; 2002 [cité 25 nov 2019]. (Chambers J, Eddy W, Härdle W, Sheather S, Tierney L. Statistics and Computing). Disponible sur: http://link.springer.com/10.1007/978-0-387-21706-2

58. Kuhn M. Caret : Classification and regression training [Internet]. 2017. Disponible sur: https://CRAN.R-project.org/package=caret

59. Guenther S-F. neuralnet: Training of Neural Networks [Internet]. 2016. Disponible sur: https://CRAN.R-project.org/package=neuralnet

60. Olden JD, Joy MK, Death RG. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecol Model. nov 2004;178(3-4):389-97.

61. Gevrey M, Dimopoulos I, Lek S. Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecol Model. févr 2003;160(3):249-64.

62. Cortes C, Vapnik V. Support-Vector Networks. Mach Learn. 1995;20(3):273-97.

63. Bzdok D, Krzywinski M, Altman N. Machine learning: supervised methods. Nat Methods. janv 2018;15(1):5-6.

64. Adeli E, Wu G, Saghafi B, An L, Shi F, Shen D. Kernel-based Joint Feature Selection and Max-Margin Classification for Early Diagnosis of Parkinson's Disease. Sci Rep. févr 2017;7(1):41069.

65. Hsu C-W, Chang C-C, Lin C-J. A practical guide to support vector classification [Internet]. 2003. Disponible sur: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/gu ide.pdf

66. Leisch D. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien [Internet]. 2017. Disponible sur: https://CRAN.R-project.org/package=e1071

67. Chatterjee S. fastAdaboost: a Fast Implementation of Adaboost [Internet]. 2016. Disponible sur: https://CRAN.R-project.org/package=fastAdaboost

68. Roy A, Perkins NJ, Buck Louis GM. Assessing Chemical Mixtures and Human Health: Use of Bayesian Belief Net Analysis. J Environ Prot. 1 juin 2012;3(6):462-8.

69. Zhao Y, Li J, Zhang M, Lu Y, Xie H, Tian Y, et al. Machine Learning Models for the Hearing Impairment Prediction in Workers Exposed to Complex Industrial Noise: A Pilot Study. Ear Hear. juin 2019;40(3):690-9.

70. Lim C, Yu B. Estimation Stability With Cross-Validation (ESCV). J Comput Graph Stat. 2 avr 2016;25(2):464-92.

71. Billionnet C, Sherrill D, Annesi-Maesano I. Estimating the Health Effects of Exposure to Multi-Pollutant Mixture. Ann Epidemiol. févr 2012;22(2):126-41.

72. Barrera-Gómez J, Agier L, Portengen L, Chadeau-Hyam M, Giorgis-Allemand L, Siroux V, et al. A systematic comparison of statistical methods to detect interactions in exposome-health associations. Environ Health. déc 2017;16(1):74.

73. Duh M-S, Walker A-M, Ayanian J-Z. Epidemiologic interpretation of artificial neural networks. Am J Epidemiol. 1998;147(12):1112-22.

74. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). Ann Stat. avr 2000;28(2):337-407.

75. Pekkanen J, Pearce N. Environmental epidemiology: challenges and opportunities. Environ Health Perspect. janv 2001;109(1):1-5.

76. Crown WH. Potential Application of Machine Learning in Health Outcomes Research and Some Statistical Cautions. Value Health. mars 2015;18(2):137-40.

77. Le Moal J, Eilstein D, Salines G. La santé environnementale est-elle l'avenir de la santé publique ? Santé Publique. 2010;22(3):281-9.

78. Tonne C, Basagaña X, Chaix B, Huynen M, Hystad P, Nawrot TS, et al. New frontiers for environmental epidemiology in a changing world. Environ Int. juill 2017;104:155-62.

XI. ANNEXES

1. Tableau A1 : Revue de la littérature

	Supervisé/ Non supervisé	Forces	Faiblesse	Application
Séléction de variables				
Deletion /Substitution/Addition (D/S/A)	Supervisé	- Moins sensible aux valeurs aberrantes - Bonne capacité prédictive	- Ces estimations ne sont pas consistantes si le ratio (taille de l'échantillon) / (nombre de variables) est petit. - Peu d'applications dans le domaine de l'épidémiologie environnementale	(Beckerman et al., 2013)
LASSO (least absolute Shrinkage and selection operator)	Supervisé	- Peut réduire les coefficients jusqu'à 0 exactement. - Donne une estimation précise grâce au compromis biais/variance	- Quand les variables sont très corrélées, ne retient qu'une seule variable parmi ce bloc et réduit les coefficients des autres variables à 0.	(Roberts & Martin, 2005) (Vriens et al., 2017)
Régression en arête (Ridge)	Supervisé	- Réduit les coefficients des variables corrélées proportionnellement vers 0.	- N'est pas capable de réduire les coefficients à 0.	(Roberts & Martin, 2005)
Elastic-Net	Supervisé	- Permets de réduire les limites du LASSO. - Autorise les données très corrélées et de haute dimension.	- Ne considère que les combinaisons linéaires et non les interactions. - Peut entraîner des biais car trop de réduction des coefficients de variables corrélées.	(Lenters et al., 2016) (Forns et al., 2016)

Réduction de la dimension				
Analyse en composantes principales (ACP)	Non supervisé	- Efficace et facile d'application - Utilisé aussi pour des données très corrélées	- Les composantes principales ne sont pas de la même unité que les variables d'exposition d'origine. - Manque de potentiel de relation entre les composantes principale et la problématique recherchée.	(Agay-Shay et al., 2015) (Berg et al., 2017) (Forns et al., 2016)
Partial least squares (les moindre carrés partiels) : PLS	Supervisé	- Plus efficace que l'ACP pour la réduction de la dimension grâce à la nature supervisée de son algorithme.	- Pas toujours optimal pour enlever les variables qui n'ont aucune association avec la variable réponse. - Etudie généralement des combinaisons linéaires entre variables.	(Berg et al., 2017)
Classification et prédiction				
Arbres de régression et de segmentation (CART)	Supervisé	- Approche non paramétrique - Facile à interpréter	- Ne tient pas directement compte des facteurs de confusion.	(Gass, Klein, Chang, Flanders, & Strickland, 2014) (Gass et al., 2015)
Random forest (ou forêts aléatoires)	Supervisé	- Approche non paramétrique - Plus stable qu'un arbre simple - Moins sujet au sur-apprentissage	- Plus difficile à interpréter qu'un arbre simple - Effet boîte noire - Ne détecte uniquement que les interactions pour des gros échantillons.	(Xu et al., 2011) (Winham et al., 2012) simulation

Boosting Trees (ou arbres boostés)	Supervisé	- Approche non paramétrique - Plus stable qu'un arbre simple - Flexible	- Plus difficile à interpréter qu'un arbre simple - Effet boîte noire	(Lampa, Lind, Lind, & Bornefalk-Hermansson, 2014)
Réseaux de neurones (ANN)	Supervisé	- Approche autorisant le non paramétrique - Largement utilisé en épidémiologie génétique et en prédiction de cancer	- Peu d'application en épidémiologie environnementale - Effet boîte noire - Les résultats dépendent de la calibration des paramètres	(Tomita et al., 2004) (Sherriff & Ott, 2004) (Saritas, 2012)
Support vector machine ou machine à vecteurs de support (SVM)	Supervisé	- Plus récent que les ANN - Peut traiter des données de haute dimension. - Non sujet au sur-apprentissage - Robuste au bruit - Analyse aussi les combinaisons non linéaires avec l'astuce du noyau	- Effet boîte noire - Calibration de plusieurs paramètres pouvant influencer les résultats - Peut-être affecté par la présence de données manquantes.	(Zheng et al., 2013) (Cuevas Tello, Hernandez-Ramirez, & Garcia-Sepulveda, 2013) (Ban, Heo, Oh, & Park, 2010)

UNIVERSITÉ DE NANTES

2. Tableau A2 : Régression logistique avec OR et p-value de chaque polluant

En gras se trouvent les polluants associés à l'endométriose de manière significative (9)

Régression Logistique Odds Ratio (OR) avec IC95% et p-value				
	OR	IC95%		p-value
Age	1,350	[0,897	2,073]	0,157
IMC	0,739	[0,479	1,110]	0,152
2.3.7.8_TCDD	1,345	[0,804	2,305]	0,266
1.2.3.7.8_PeCDD	**1,796**	**[1,055**	**3,211]**	**0,037**
1.2.3.4.7.8_HxCDD	1,363	[0,812	2,346]	0,248
1.2.3.6.7.8_HxCDD	1,325	[0,807	2,222]	0,272
1.2.3.7.8.9_HxCDD	1,083	[0,715	1,663]	0,707
1.2.3.4.6.7.8_HpCDD	0,914	[0,594	1,391]	0,675
OCDD	0,986	[0,651	1,493]	0,947
2.3.7.8_TCDF	0,822	[0,532	1,253]	0,364
1.2.3.7.8_PeCDF	0,919	[0,604	1,392]	0,688
2.3.4.7.8_PeCDF	1,269	[0,757	2,179]	0,371
1.2.3.4.7.8_HxCDF	1,115	[0,724	1,744]	0,623
1.2.3.6.7.8_HxCDF	1,041	[0,669	1,629]	0,859
1.2.3.7.8.9_HxCDF	1,263	[0,835	1,966]	0,280
2.3.4.6.7.8_HxCDF	0,947	[0,614	1,456]	0,801
1.2.3.4.6.7.8_HpCDF	0,931	[0,618	1,398]	0,727
1.2.3.4.7.8.9_HpCDF	1,233	[0,819	1,900]	0,323
OCDF	**4,596**	**[2,470**	**9,839]**	**0,000**
PCB77	**0,582**	**[0,366**	**0,890]**	**0,016**
PCB81	0,726	[0,464	1,107]	0,145
PCB126	1,175	[0,726	1,929]	0,512
PCB169	1,135	[0,614	2,124]	0,685
PCB105	**1,741**	**[1,077**	**2,935]**	**0,029**
PCB114	1,826	[1,020	3,457]	0,051
PCB118	**1,976**	**[1,170**	**3,519]**	**0,014**

PCB123	**1,993**	**[1,205**	**3,470]**	**0,010**
PCB156	0,999	[0,511	1,941]	0,998

Régression Logistique Odds Ratio (OR) avec IC95% et p-value				
	OR	IC95%		p-value
PCB157	0,781	[0,414	1,427]	0,427
PCB167	1,478	[0,816	2,768]	0,205
PCB189	0,810	[0,385	1,661]	0,567
PCB28	0,758	[0,484	1,157]	0,209
PCB52	1,208	[0,801	1,860]	0,375
PCB101	1,148	[0,758	1,756]	0,516
PCB138	1,445	[0,835	2,585]	0,196
PCB153	1,367	[0,758	2,540]	0,304
PCB180	0,882	[0,436	1,750]	0,719
PBDE28	1,149	[0,758	1,793]	0,521
PBDE47	1,198	[0,788	1,873]	0,408
PBDE99	0,968	[0,638	1,476]	0,880
PBDE100	1,102	[0,725	1,702]	0,651
PBDE153	1,077	[0,696	1,680]	0,739
PBDE154	0,838	[0,550	1,263]	0,401
PBDE183	1,448	[0,951	2,289]	0,095
PBDE209	1,267	[0,835	2,031]	0,283
PBB153	**4,106**	**[1,794**	**10,886]**	**0,002**
a.HBCD	0,895	[0,585	1,354]	0,598
HCB	**1,700**	**[1,071**	**2,852]**	**0,032**
beta_HCH	1,617	[1,007	2,706]	0,053
trans_nonachlore	**1,920**	**[1,155**	**3,394]**	**0,017**
oxychlordane	**2,736**	**[1,484**	**5,729]**	**0,003**
cis heptachlore epoxyde	**3,227**	**[1,813**	**6,510]**	**0,000**
dieldrine	**2,310**	**[1,441**	**3,912]**	**0,001**
p_p_DDT	1,062	[0,705	1,617]	0,773
p_p_DDE	0,901	[0,588	1,359]	0,618

UNIVERSITÉ DE NANTES

3. Tableau A3 : Coefficients obtenus avec la régression logistique pénalisée Elastic-Net (29)

Polluants	Coefficient de régression	Polluants	Coefficient de régression
OCDF	4,443	**PCB 126**	0
Cis heptachlore epoxyde	1,317	**PCB 105**	0
1.2.3.7.8 PeCDD	1,172	**PCB 114**	0
PCB 123	1,131	**PCB 138**	0
PCB 101	1,087	**PCB 153**	0
PBDE 183	0,963	**PBDE 99**	0
PCB 52	0,626	**dieldrine**	0
PCB 169	0,621	**IMC**	-0,041
PBB 153	0,515	**HCB**	-0,083
PBDE 28	0,497	**P DDE**	-0,111
1.2.3.6.7.8 HxCDD	0,476	**PCB 28**	-0,149
1.2.3.7.8.9 HxCDF	0,463	**PBDE 100**	-0,211
PBDE 47	0,434	**PCB 167**	-0,235
oxychlordane	0,414	**PCB 156**	-0,246
OCDD	0,395	**PBDE 209**	-0,320
P DDT	0,292	**1.2.3.7.8.9 HxCDD**	-0,320
1.2.3.4.7.8 HxCDD	0,251	**2.3.7.8 TCDD**	-0,329
β HCH	0,114	**PCB 157**	-0,330
PCB 118	0,090	**1.2.3.4.6.7.8 HpCDF**	-0,434
Age	0,040	**PCB 180**	-0,489
Trans nonachlore	0,032	**α HBCD**	-0,517
1.2.3.7.8 PeCDF	0,009	**PCB 189**	-0,721
1.2.3.4.7.8 HxCDF	0,005	**1.2.3.4.7.8.9 HpCDF**	-0,727
2.3.7.8 TCDF	0	**1.2.3.4.6.7.8 HpCDD**	-0,811
2.3.4.7.8 PeCDF	0	**PBDE 154**	-0,908
1.2.3.6.7.8 HxCDF	0		

2.3.4.6.7.8 HxCDF	-1,067
PCB 81	-1,278
PBDE 153	-1,319
PCB 77	-1,738

4. Tableau A4 : Polluants de la base de données ENDOTOX

Dioxines : polychloro-dibenzo-para-dioxines
2.3.7.8 - TCDD
1.2.3.7.8 - PeCDD
1.2.3.4.7.8 - HxCDD
1.2.3.6.7.8 - HxCDD
1.2.3.7.8.9 - HxCDD
1.2.3.4.6.7.8- HpCDD
OCDD
Furanes : polychloro-dibenzo-furanes
2.3.7.8 - TCDF
1.2.3.7.8 - PeCDF
2.3.4.7.8 - PeCDF
1.2.3.4.7.8 - HxCDF
1.2.3.6.7.8 - HxCDF
1.2.3.7.8.9 - HxCDF
2.3.4.6.7.8 - HxCDF
1.2.3.4.6.7.8 -HpCDF
1.2.3.4.7.8.9 -HpCDF
OCDF
Polychloro biphényles (PCB)
PCB dioxine-like (DL)
PCB 77
PCB 81
PCB 126
PCB 169
PCB 105
PCB 114
PCB 118

PCB123
PCB 156
PCB 157
PCB 167
PCB 189
PCB non dioxine-like (NDL)
PCB 28
PCB 52
PCB 101
PCB 138
PCB 153
PCB 180
Retardateurs de flammes
Polybromodiphényléthers (PBDE)
PBDE 28
PBDE 47
PBDE 99
PBDE 100
PBDE 153
PBDE 154
PBDE 183
PBDE209
Polybromobiphényles (PBB)
PBB 153
hexabromocyclododecane (HBCD)
α HBCD
Pesticides organochlorés
HCB = Hexachlorobenzène
β-HCH = Hexachlorocyclohexane
trans-nonachlore
oxychlordane
cis-heptachlore epoxyde
dieldrine
p,p'-DDT = p,p’-Dichlorodiphényltrichloroéthane
p,p'-DDE = p,p’-Dichlorodiphénylether

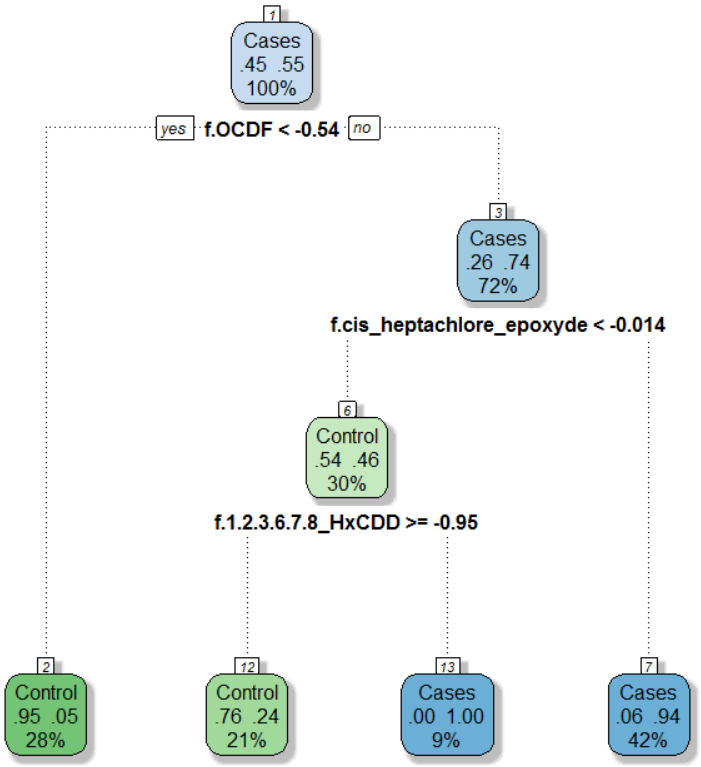
5. Tableau A5 : Liste des packages et fonctions utilisées dans le logiciel R

Package	Fonction
Préparation des données	
« base »	« log » , « scale »
« mice »	« mice »
Application des algorithmes	
Séparation du jeu de données, validation croisée et entraînement de l'algorithme	
« caret »	« createDataPartition », « trainControl », « train »
ANN	
« nnet »	« nnet »
« neuralnet »	« neuralnet »
« NeuralNetTools » (visualisation des données)	« plotnet »
SVM	
« e1071 »	« svm »
« kernlab » (fonction noyau des SVM)	« ksvm » (Noyau à base radiale)
Arbres « boostés »	
« fastAdaboost »	« Adaboost.M1 »
Performance de prédiction	
« caret »	« predict » (Prédiction de l'algorithme sur jeu test)
« caret »	« confusionMatrix » (Générer la matrice de confusion)
« pROC »	« roc » et « auc » (pour obtenir la courbe ROC et l'AUC)

6. Interprétation d'un arbre de classification simple avec la base ENDOTOX

Avant de commenter les résultats du « boosting tree », nous avons implémenté un arbre simple pour comprendre comme celui-ci s'interprète avant d'en faire les itérations. Les paramètres utilisés ici sont ceux par défaut, car le modèle nous servira surtout à illustrer graphiquement un arbre simple (ici le paramètre se nomme le paramètre de complexité : cp, qui est à 0 par défaut).

Concernant les résultats de l'arbre décisionnel simple : l'interprétation est assez facilement compréhensible. Chaque nœud de l'arbre représente une variable explicative avec un seuil, et à l'intérieur sera précisé la proportion des témoins (pas d'endométriose) et des cas (endométriose). A la fin, dans les nœuds terminaux (ou feuille) l'objectif est d'obtenir des nœuds très homogènes, c'est-à-dire soit ne contenant que des cas, soit que des témoins, pour pouvoir à la fin bien identifier les cas des témoins.



Interprétation des résultats d'un arbre simple représentés par le schéma ci-dessous :

Le premier nœud représente la totalité du jeu d'entrainement (80% de la totalité du jeu de données), c'est-à-dire 80 femmes, avec 36 témoins et 44 cas (45% et 55%). La première variable identifiée, permettant de séparer ce jeu de données est la variable OCD Si sa concentration est inférieure à -0.54 pg.gl.w, ce qui est le cas ici pour 28% des femmes, et bien nous avons 95% de témoins : c'est-à-dire que pour une concentration très faible en OCDF, les femmes n'ont généralement pas l'endométriose. En revanche, si leur concentration d'OCDF est supérieure à -0.54 pg.gl.w (ce qui est le cas pour 72% des femmes, i-e 58 femmes), on retrouve une plus grande proportion de cas (74%, ou 43 femmes) mais il reste 26% de témoins (15 femmes). L'algorithme continue donc la séparation pour obtenir des nœuds le plus pur possible (100% de témoins ou 100% de cas). Une nouvelle dichotomie est alors lancée avec la 2ème variable : le cis heptachlore époxyde (avec un seuil pris à -0,014 pg.gl.w), puis si celle-ci est < à -0,014 pg.gl.w, une troisième séparation s'opère avec la variable : la dioxine 1.2.3.6.7.8 HxCDF avec un seuil choisi à 0,95 pg.gl.w. Nous obtenons 4 nœuds terminaux, et globalement l'interprétation qu'il en ressort est que les cas se retrouvent plus souvent avec une concentration d'OCDF élevée et une concentration de cis heptachlore élevée ou alors avec une concentration en OCDF élevée mais une faible en cis heptachlore et faible en dioxine HxCDD. Les témoins se retrouvent principalement dans ceux ayant une concentration faible en OCDF ou alors plus élevée en OCDF faible en cis heptachlore mais plus élevée en dioxine HxCDD.

Les 3 variables les plus importantes de ce modèle sont donc OCDF, cis heptachlore époxyde et la dioxine 1.2.3.6.7.8 HxCDD.

7. Tableau A6 : Résultats coefficients (poids) des Réseaux de neurones

Ici, on peut voir le coefficient de chaque variable issu du réseau de neurone via la formule d'Olden (Section III,3,a,iii) : le résultat est la somme des produits des poids synaptiques de la couche d'entrée à la couche de sortie.

Variables explicatives	Poids synaptiques
OCDF	5.758
PCB 77	3.176
Cis heptachlore époxyde	2.500
PBB 153	1.954
PCB 123	1.728
PBDE 153	1.686
1.2.3.7.8 PeCDD	1.633
PBDE 154	1.524
2.3.4.6.7.8 HxCDF	1.451
PCB 52	1.380
PCB 101	1.327
α HBCD	1.291
IMC	1.275
PCB 126	1.221
PCB 28	1.173
dieldrine	1.111
1.2.3.4.6.7.8 HpCDD	1.053
PBDE 183	1.050
oxychlordane	1.038
PBDE 47	1.009

Ici, on peut voir le coefficient de chaque variable issu du réseau de neurone via la formule d'Olden : le résultat est la somme des produits des poids synaptiques de la couche d'entrée à la couche de sortie.

UNIVERSITÉ DE NANTES

XII. RÉSUMÉ

1. Résumé en français (500 mots)

Les facteurs environnementaux contribuent à l'étiologie d'un grand nombre de pathologies. Les études de bio-surveillance montrent que nous sommes exposés à une multitude de produits chimiques, mais la caractérisation de l'impact sur la santé de ces polluants en mélange complexe reste un défi majeur. Une des limites des approches statistiques classiques usuellement mises en œuvre en épidémiologie environnementale, de type régression linéaire et régression logistique, est de ne pas tenir compte de ce « cocktail » de polluants et de leur interactions, ce qui devrait être au contraire considéré dans des scénarios réalistes. La fiabilité de ces modèles statistiques linéaires se trouve altérée, notamment en raison de la multicollinéarité et de la redondance des variables chimiques en épidémiologie environnementale. Il existe aujourd'hui un nombre croissant de méthodes alternatives dans le domaine de l'apprentissage automatique, pour pouvoir pallier à ces contraintes. Dans le cadre de ce travail, on a considéré l'endométriose, une pathologie hormono-dépendante dont l'étiologie environnementale a déjà été démontrée depuis quelques années par des modèles statistiques linéaires. Dans ce travail nous avons étudié l'association entre les niveaux d'exposition internes de polluants (biomarqueurs d'exposition mesurés dans le tissu adipeux) et la présence d'endométriose dans une étude cas-témoin de 99 femmes des Pays de la Loire (France), via trois approches « multi-polluants » d'apprentissage automatique, les réseaux de neurones, les machines à vecteurs de support, et les « Boosting Trees ». Ces trois approches ont été sélectionnées pour leur originalité dans le domaine de l'épidémiologie environnementale, mais aussi par le fait qu'elles puissent intégrer un grand nombre de co-variables et les évaluer en tenant compte de leurs éventuelles synergies. Les résultats montrent que la méthode la plus performante des trois identifiée ici, en termes de capacité de prédiction, sont les réseaux de neurones, suivi des « Boosting Trees » puis en dernière, mais restant toujours performante, les machines à vecteurs de support. Ces trois approches ont permis de mettre en évidence, en commun, quatre polluants fortement associés à l'endométriose : l'OCDF, le cis-heptacholre-époxyde, le PCB 123, le PBB 153 et le 1.2.3.7.8 PeCDD. Ces polluants sont également mis en évidence en utilisant des modèles linéaires tels que la régression logistique, non pénalisée ou avec une pénalisation de type Elastic-Net. Les approches d'apprentissage automatique testées semblent ainsi fournir des résultats cohérents entre elles. De plus, celles-ci paraissent potentiellement aptes à gérer un enjeu majeur en épidémiologie environnementale qui est de prendre en compte des mélanges de polluants pour se rapprocher de scénarios d'exposition réalistes. Cependant, ces méthodes restent assez complexes à mettre en œuvre et à paramétrer, et leurs résultats restent en partie difficilement interprétables. Des travaux complémentaires paraissent donc nécessaires afin d'approfondir les connaissances théoriques relatives à ces approches, notamment sur le plan biostatistique et bioinformatique, afin de réduire leur effet « boîte noire », de mieux assoir le choix éclairé d'une approche méthodologique préférentielle, de pouvoir les calibrer de manière optimale, et

de pouvoir aboutir à une interprétation plus immédiatement intelligible et compréhensible sur le plan pratique.

2. Summary in english (500 words)

The current state of literature, informs us that a large number of pathologies may be associated to environmental risk factors. Biomonitoring studies have shown that humans are exposed to many mixtures of chemical products during our lifetimes which can be potentially toxic. A major limitation to our understanding is the use of conventional statistical approaches in environmental epidemiology like linear regression or logistic regression, which are unable to analyse this “cocktail” of highly correlated exposures. Consequently, these conventional approaches analyse only a single chemical exposure, one at a time, adjusting for confounding factors. This is hardly reflective of a realistic scenario. Additionally, data on environmental exposures are highly intercorrelated and redundant (with similar chemical properties), which can affect the reliability of these conventional methods and yield high variability in the coefficients. Recent literature reflects a growing interest in alternative methods of statistical analyses such as “machine learning” algorithms, each with their own strengths and weaknesses, which seek to overcome these aforementioned issues and assess the health effect of pollutant mixtures.

This project focuses on the link between a mixture of endocrine disrupting chemicals and endometriosis, a gynecological and estrogen-dependent disease, whose environmental etiology has already been supported over the past years with linear statistical approaches. It is thus necessary to assess the health impact of these mixtures of environmental endocrine disrupting pollutants on endometriosis. Here, we studied associations between internal levels of pollutants present in adipose tissue (biomarkers of exposure) and the presence of endometriosis in a case-control study of 99 women from Pays de la Loire, France, with three “multi-pollutant” models of machine learning: Neural Networks, Support Vector Machines and Boosting Trees. These three algorithms were selected for their applicability in the field of environmental epidemiology and for their ability to incorporate a large number of interrelated covariables which impact endometriosis. Additionally, these algorithms are also able to overcome highly intercorrelated data and deal with non-linear relationships between variables. The results show that neural network has the best predictive performance, followed by boosting trees and lastly but still performant : the support vector machine.

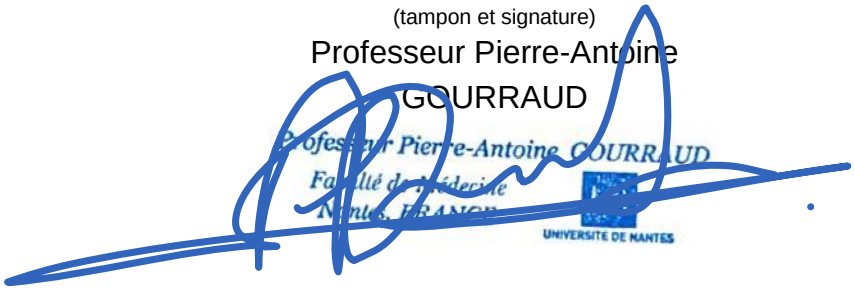
These three algorithms identified from the dataset four pollutants strongly associated with endometriosis: octachlorodibenzofuran (OCDF), cis-heptachlor epoxide, polychlorinated biphenyl (PCB) 123, the polybrominated biphenyl (PBB) 153 and the 1.2.3.7.8 pentachlorodibenzo*para*dioxine (PeCDD). These findings are consistent with the results from the preliminary study, wherein these four chemicals were also identified by the traditional linear models (i.e. non-penalized logistic regression and penalized logistic regression, Elastic-Net). These new statistical approaches thus present a promising development in environmental epidemiology because of their ability to account

for pollutant mixtures and more accurately represent realistic scenarios. However, these methods are still poorly understood and complicated to implement, and their results are difficult to interpret. Future work will be required to improve biostatistical, mathematical and bioinformatic theoretical knowledge of these models in order to reduce their "black box" effect, to better guide model selection, to optimize calibration techniques, and to interpret the results with relevance to human epidemiology.

UNIVERSITÉ DE NANTES

XIII. SIGNATURES

Vu, le Président du Jury,
(tampon et signature)
Professeur Pierre-Antoine GOURRAUD



Vu, le Directeur de Thèse,
(tampon et signature)
Dr German CANO-SANCHO

Vu, le Doyen de la Faculté,
Professeur Pascale JOLLIET

UNIVERSITÉ DE NANTES

NOM : MOURET **PRENOM** : Delphine

TITRE DE THESE

Etude de l'association entre l'exposition aux polluants organiques persistants (POPs) et l'endométriose, à l'aide d'algorithmes d'apprentissage statistique.
Exploitation des données du projet ENDOTOX : une étude cas-témoins réalisée en Pays de la Loire entre 2013 et 2015

Study of the association between exposure to persistent organic pollutants (POPs) and endometriosis, using statistical learning algorithms.
Use of ENDOTOX project data: a case-control study conducted in Pays de la Loire between 2013 and 2015

RESUME

Dans ce travail nous avons étudié l'association entre les niveaux d'exposition internes de polluants et la présence d'endométriose dans une étude cas-témoin de 99 femmes des Pays de la Loire (France) entre 2013 et 2015, via trois approches « multi-polluants » d'apprentissage automatique, les réseaux de neurones, les machines à vecteurs de support, et les « Boosting Trees ». Les résultats de cette étude montrent que plusieurs polluants sont associés à l'endométriose. Ces trois approches peuvent intégrer un grand nombre de co-variables et les évaluer en tenant compte de leurs éventuelles synergies et de leur correlation, là où les modèles traditionnels de régression linéraire et logistique ont montré leurs limites. Elles semblent donc permettre de répondre à un enjeu majeur en épidémiologie environnementale qui est de prendre en compte des mélanges de polluants pour se rapprocher de scénarios d'exposition réalistes. Cependant, ces méthodes restent assez complexes à mettre en œuvre, et leurs résultats restent en partie difficilement interprétables. Des travaux complémentaires paraissent donc nécessaires afin d'approfondir leurs ces méthodes qui pourraient également être intégrées dans la formation des épidémiologistes voire des professionnels de santé afin de répondre aux nouveaux enjeux de la médecine et du traitement des données de santé.

We studied the association between the internal exposure levels of pollutants and the presence of endometriosis in a case-control study of 99 women from Pays de la Loire (France) between 2013 and 2015, with three machine learning approaches : neural networks, support vector machines, and "Boosting Trees". The results of this study show that several pollutants are associated with endometriosis. These three approaches can integrate a large number of co-variables and evaluate them taking into account their possible synergies and their correlation, where the traditional linear and logistic regression models have shown their limits. They seem to be able to respond to a major challenge in environmental epidemiology which is to take into account mixtures of pollutants to get closer to realistic exposure scenarios. However, these methods are rather complex to implement, and their results remain partly difficult to interpret. Further work is needed to better know these methods, which could also be integrated into the training of epidemiologists or even health professionals to answer to the new challenges of medicine and the processing of health data.

MOTS-CLES

Endométriose ; Polluants organiques persistants ; Épidémiologie environnementale ; Etude observationnelle ; Méthodes d'apprentissage automatique.
Endometriosis ; Persistent organic pollutants ; Environmental epidemiology ; Observational study ; statistical learning methods.