Nantes Université

IAE Nantes
Économie & Management
Pôle Sociétés
Nantes Université

Mémoire

Master 1 Econométrie et Statistique

Parcours Econométrie Appliquée

Intelligence artificielle et santé : machine learning appliqué au cancer du sein

DUPAS-BROUSSE Simon

Directeur de master : **DARNE Olivier**

Professeur encadrant : **SUIRE Raphaël**

Juin 2024

Remerciements

Je souhaite tout d'abord exprimer ma profonde reconnaissance à mon tuteur de mémoire, Mr Raphaël Suire, professeur à l'IAE de Nantes. Son expertise et sa connaissance approfondie ont été des éléments clés dans la réalisation de ce projet. Ses conseils avisés et son orientation précise ont grandement facilité ce travail. Je tiens également à remercier chaleureusement toutes les personnes qui, de près ou de loin, m'ont soutenu et encouragé tout au long de mon parcours académique.

Résumé

Cette étude explore l'utilisation et l'adaptabilité de l'intelligence artificielle aux données de santé, en se concentrant sur les modèles de machine learning appliqués aux caractéristiques cellulaires des tumeurs du cancer du sein. La problématique centrale est d'identifier les modèles prédictifs les plus efficaces pour classifier les tumeurs en fonction de leur malignité, en utilisant un ensemble de données historiques du Dr William H. Wolberg datant de 1992. La méthodologie inclut le prétraitement des données, le suréchantillonnage pour équilibrer les classes, et l'évaluation des modèles à l'aide de métriques de performance telles que la sensibilité, la spécificité, et la précision. Les résultats montrent que la régression logistique a obtenu la meilleure performance (96.3%), suivie par les réseaux de neurones (95.56%) et les arbres de décision (93.33%). Ces résultats illustrent que les modèles plus simples peuvent parfois surpasser les modèles plus complexes en termes de performance prédictive.

Abstract

This study explores the use and adaptability of artificial intelligence in healthcare data, focusing on machine learning models applied to the cellular characteristics of breast cancer tumors. The central issue is to identify the most effective predictive models for classifying tumors based on their malignancy, using a historical dataset from Dr. William H. Wolberg dating back to 1992. The methodology includes data preprocessing, oversampling to balance classes, and model evaluation using performance metrics such as sensitivity, specificity, and precision. The results show that logistic regression achieved the best performance (96.3%), followed by neural networks (95.56%) and decision trees (93.33%). These findings illustrate that simpler models can sometimes outperform more complex models in terms of predictive performance.

Mots clés : *Intelligence artificielle, Données de santé, Machine learning, Cancer du sein*

Sommaire

1. Introduction

L'intégration de l'intelligence artificielle (IA) et du Machine Learning (ML) dans le domaine médical a une dimension historique fascinante qui témoigne de l'évolution rapide de la technologie et de la médecine. Les premières applications remontent aux années 1960, lorsque des chercheurs ont commencé à explorer les possibilités de l'IA pour aider à interpréter les résultats d'examens médicaux. Cependant, ce n'est que dans les décennies suivantes que les progrès technologiques ont permis des avancées significatives. Les années 1980 et 1990 ont vu l'émergence de systèmes experts capables de diagnostiquer des maladies spécifiques en analysant des symptômes et des données médicales. Avec l'arrivée d'algorithmes d'apprentissage automatique plus avancés et de la disponibilité de grandes quantités de données médicales, les années 2000 ont marqué une nouvelle ère pour l'IA en médecine.

L'intelligence artificielle (IA) est un domaine de l'informatique qui se concentre sur la création de systèmes informatiques capables d'effectuer des tâches qui nécessitent généralement une intelligence humaine. L'objectif de l'IA est de développer des algorithmes et des techniques qui permettent aux machines d'imiter les capacités cognitives humaines et d'accomplir des tâches de manière autonome.

Le Machine Learning (ML), est une branche de l'IA qui se concentre sur le développement de techniques permettant aux ordinateurs d'apprendre à partir de données et d'améliorer leurs performances sur des tâches spécifiques sans être explicitement programmés pour chaque cas. Au lieu de suivre des instructions programmées, les systèmes de ML utilisent des algorithmes qui analysent des données pour identifier des modèles et tirer des conclusions. Ces modèles peuvent être utilisés pour effectuer des prédictions ou des décisions sur de nouvelles données.

Ces dernières années, l'avènement de ChatGPT et plus généralement des Large Language Models (LLM) a suscité un regain d'intérêt pour l'intelligence artificielle (IA). En effet, il n'a pas fallu longtemps avant de voir apparaître les premières applications concrètes dans le domaine de la santé. En 2023, une IA conversationnelle développée

par Nuance (filiale de Microsoft) a permis aux professionnels de santé d'utiliser leur voix au lieu de leur clavier pendant les consultations, « réduisant ainsi de 30% le temps passé à documenter chaque consultation, et récupérant ainsi 2 minutes par patient. En fin de compte, le gain de temps est considérable – de l'ordre d'une heure par médecin – et nous permet de soigner plus de patients », déclare Pieter Nel, médecin urgentiste à l'hôpital de Mackay (Australie).[i]

Cependant, les tâches administratives ne sont qu'une partie du domaine d'application de l'IA en médecine. L'un des domaines les plus courants est celui de la médecine prédictive, où elle aide les professionnels de santé dans leurs diagnostics.

À cette fin, l'étude Ng, A.Y., Oberije, C.J.G., Ambrózay, É. et al.[ii] évalue un système d'IA d'imagerie médicale. Ce dernier, disponible dans le commerce, est mis en œuvre en tant que lecteur supplémentaire à la double lecture standard ayant pour but d'améliorer le dépistage du cancer du sein. Les résultats plutôt favorables, suggèrent une amélioration de la détection précoce du cancer du sein grâce à ce système d'IA. Néanmoins, l'étude souligne qu'un seul système d'IA commercial a été évalué et que les résultats peuvent ne pas être représentatifs d'autres systèmes disponibles dans le commerce.

En effet, malgré des avancées prometteuses, il existe plusieurs limites à prendre en compte. L'une des principales réside dans le défi de sélectionner les modèles d'IA les plus pertinents parmi une multitude d'options disponibles. Avec la prolifération des techniques d'apprentissage automatique et des algorithmes d'IA, il peut être difficile pour les professionnels de santé de faire le tri entre les différents modèles et de déterminer lesquels sont les plus adaptés à leurs besoins spécifiques. Cette tâche exige une expertise approfondie dans le domaine de l'IA ainsi qu'une compréhension des exigences cliniques et des données médicales pertinentes.

Etant donné l'importance primordiale de la santé et les défis démographiques auxquels les générations futures seront confrontées, il devient essentiel d'explorer les avancées rendues possibles par une technologie en plein essor. Une compréhension approfondie des mécanismes sous-jacents à la création de modèles prédictifs orientera de manière optimale leur utilisation par les acteurs de la santé.

Dans cette optique, ce mémoire se focalise sur l'analyse de différents modèles de Machine Learning appliqués au cancer du sein. Ce cancer présente un intérêt particulier en raison de sa prévalence élevée et de son impact significatif sur la santé des individus, en particulier des femmes.

En examinant la manière dont les modèles prédictifs peuvent être élaborés et appliqués dans le contexte particulier du cancer du sein, l'objectif est de fournir une compréhension approfondie de ces modèles afin de faciliter leur utilisation pour qu'elle soit optimale.

Ce mémoire est structuré en trois parties distinctes. La première partie se concentre sur l'étude des principales caractéristiques du cancer du sein. La seconde partie aborde les divers modèles de Machine Learning (ML) utilisés pour la classification. Enfin, la dernière partie met en pratique ces modèles en les appliquant empiriquement à une base de données dédiée au cancer du sein.

2. Le cancer du sein

2.1 Historique et Mécanismes

Premièrement, le cancer a été une maladie connue depuis l'Antiquité, caractérisée par une croissance cellulaire anarchique et incontrôlée. Cette maladie complexe résulte de mutations génétiques qui perturbent les mécanismes de régulation normaux des cellules. En conséquence, les cellules cancéreuses se multiplient de manière incontrôlée, formant des tumeurs malignes qui peuvent envahir les tissus environnants et se propager à d'autres parties du corps, un processus appelé métastases. Le cancer peut affecter pratiquement tous les tissus et organes du corps, donnant lieu à une grande variété de types de cancer.

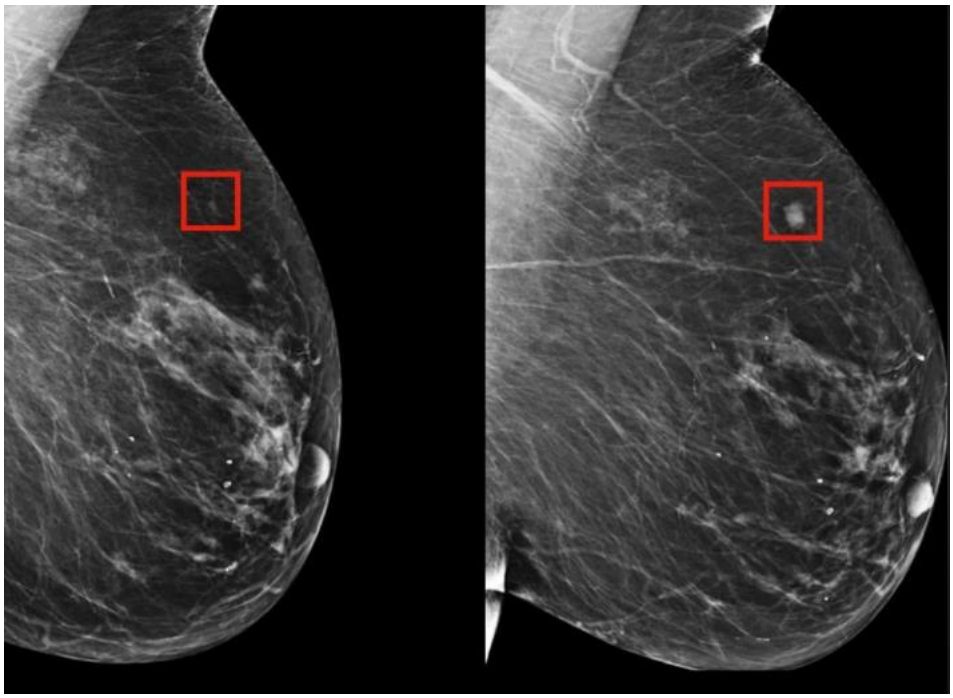
En ce qui concerne le cancer du sein, il s'agit d'une forme spécifique de cancer qui se développe dans les tissus mammaires. Cette maladie peut survenir chez les hommes mais elle est beaucoup plus fréquente chez les femmes.

La découverte et le diagnostic du cancer du sein remontent à plusieurs siècles. Les premières références historiques au cancer du sein remontent à l'Antiquité, où des observations de tumeurs mammaires ont été documentées.[iii] Cependant, les connaissances médicales et les moyens de diagnostic étaient limités à cette époque, et le cancer du sein était souvent confondu avec d'autres affections mammaires.

Ce n'est qu'au cours des derniers siècles que des progrès significatifs ont été réalisés dans le diagnostic et le traitement du cancer du sein. Au 19ème siècle, des avancées dans le domaine de la pathologie ont permis une meilleure compréhension de la nature du cancer, notamment la distinction entre tumeurs bénignes et malignes. Les premières techniques chirurgicales pour traiter le cancer du sein ont également été développées à cette époque, bien que souvent rudimentaires et associées à un taux de survie relativement faible.

Au fil du temps, des progrès ont été réalisés dans les techniques de diagnostic, notamment l'introduction de la mammographie dans la seconde moitié du 20ème siècle, qui permet de détecter les anomalies mammaires avant même qu'elles ne deviennent palpables. Des méthodes d'imagerie avancées telles que l'échographie et l'IRM ont également contribué à améliorer le diagnostic précoce du cancer du sein.

Figure 1 : Détection précoce d'une tumeur par Mammographie[1]



Nous savons que le cancer du sein est à la fois le plus fréquent et le plus meurtrier des cancers chez la femme. « En 2012, on estimait le nombre de nouveaux cas de ce cancer en France à 48 800 et le nombre de décès à 11 900 »[iv]. Devant ces chiffres préoccupants, une question légitime à se poser est : pourquoi la croissance cellulaire incontrôlée se produit-elle particulièrement, ou plus fréquemment, dans le sein ?

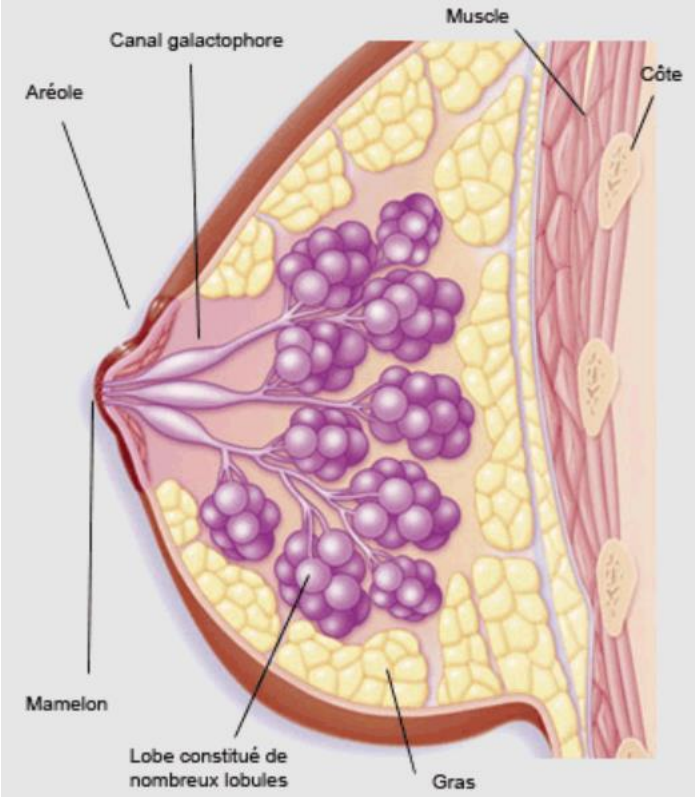
Il existe plusieurs raisons à cela, la première est la stimulation hormonale. En effet, le sein est un organe sensible aux hormones, en particulier aux œstrogènes et à la progestérone. Ces hormones sont impliquées dans la régulation de la croissance et du

[1] https://interinfos.net/wp-content/uploads/2020/01/Cancer-du-sein-2.jpg

développement des cellules mammaires. Des fluctuations hormonales, telles que celles qui surviennent pendant la menstruation, la grossesse ou la ménopause, peuvent influencer la croissance cellulaire dans le sein. Des déséquilibres hormonaux peuvent favoriser la croissance cellulaire incontrôlée.

Une autre raison est sa structure. Le sein est un organe complexe composé de nombreux types de tissus, y compris des glandes, des canaux, du tissu conjonctif et du tissu adipeux. Les canaux et les lobules, où apparaissent respectivement les carcinomes canalaires et lobulaires[2], sont particulièrement sujets au développement de tumeurs. Cette complexité structurelle rend le sein plus sujet aux erreurs dans la régulation de la croissance cellulaire.

Figure 2 : Organisation du complexe mammaire



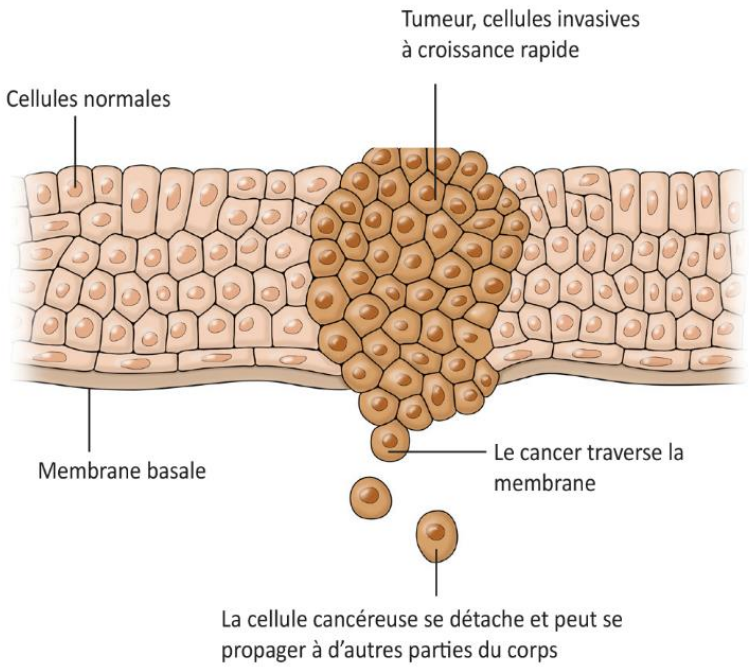
[2] Ce sont les deux types de cancer du sein les plus fréquents.

Pour finir, des mutations génétiques, telles que celles des gènes BRCA1 et BRCA2, peuvent augmenter considérablement le risque de cancer du sein. Cependant, ces mutations génétiques héréditaire ne représentent que « 5 à 10 % des cancers du sein ».[v]

Ces facteurs contribuant à la prévalence élevée ont stimulé de nombreux progrès dans le diagnostic du cancer du sein au fil des années.

En rentrant un peu plus dans les détails d'un cancer « in situ »[3], nous constatons que lorsque les cellules sont confinées dans un espace restreint, la membrane nucléaire de leur noyau est soumise à des contraintes mécaniques importantes et peut finir par se rompre. Cela entraîne la dispersion de l'ADN dans le cytoplasme, où il se dégrade. Les chercheurs ont observé que les tumeurs mammaires, qui se forment à l'intérieur des canaux galactophores, subissent également ces contraintes mécaniques au cours de leur croissance. Elles augmentent de taille, déformant les canaux et finissant par détruire leur enveloppe pour envahir les tissus environnants. [vi]

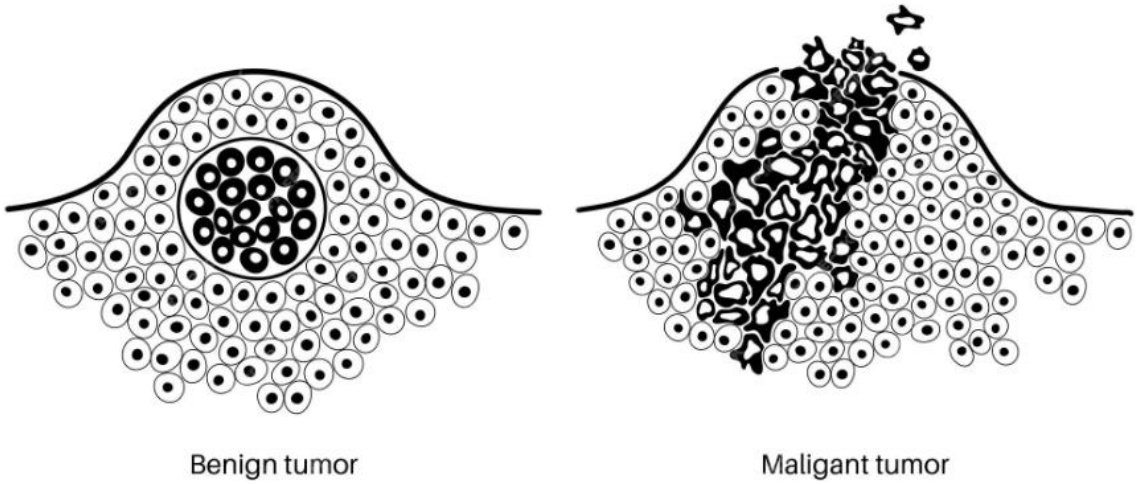
Figure 3 : Propagation de cellules cancéreuses



[3] Se trouvant dans la région du complexe mammaire et n'ayant pas encore fait de métastases vers d'autres organes.

L'approche la plus couramment utilisée pour déterminer si une tumeur observée est bénigne ou maligne est l'analyse histopathologique des échantillons de tissus prélevés lors d'une biopsie. Cette méthode permet à un pathologiste de visualiser les cellules au microscope et d'identifier des caractéristiques morphologiques distinctes entre les cellules bénignes et malignes.

Figure 4 : Tumeur bénigne et maligne



Une tumeur bénigne du sein est une masse non cancéreuse qui ne se propage pas à d'autres parties du corps, contrairement aux tumeurs malignes qui prolifèrent de manière anarchique dans un tissu ou un organe et peuvent s'étendre à d'autres parties du corps en fabriquant des métastases, conférant au cancer du sein son potentiel létal. La tumeur bénigne a généralement une forme régulière et lisse et ne menace pas les organes vitaux, ses conséquences sont principalement esthétiques ou douloureuses. Une fois traitée, les récidives de tumeurs bénignes sont rares.[vii]

Nous utiliserons ces caractéristiques cellulaires comme variables dans les modèles de classification pour déterminer la catégorie de chaque tumeur observée.

2.2 Définitions des variables

Dans cette partie, nous établirons un lien entre nos variables, définies par les caractéristiques cellulaires permettant de distinguer les tumeurs bénignes des tumeurs malignes, et les recherches déjà existantes dans la littérature.

2.2.1 Clump_thickness

Dans les diagnostics de cancer, particulièrement dans les analyses cytologiques (comme celles utilisées pour diagnostiquer le cancer du sein via des biopsies ou des frottis), le terme "clump thickness" ou "épaisseur de l'amas" est une mesure utilisée pour évaluer la densité et la disposition des cellules dans un échantillon de tissu.[viii]

L'épaisseur de l'amas peut donner des indices sur la nature des cellules. Par exemple, des amas épais de cellules peuvent être plus suggestifs de malignité, tandis que des amas plus fins pourraient indiquer une tumeur bénigne ou un tissu normal.

2.2.2 Size_uniformity

En ce qui concerne l'uniformité de la taille des cellules, les cellules cancéreuses tendent à présenter une variation significative en taille, en raison de leur croissance rapide et des altérations génétiques. Une grande variation dans la taille des cellules, observée en conjonction avec d'autres caractéristiques cytologiques, peut suggérer la présence d'une tumeur maligne, guidant ainsi le diagnostic et le traitement.

2.2.3 Shape_uniformity

Lors de l'analyse cytologique des tissus tumoraux, l'uniformité de la forme des cellules est une caractéristique observée conjointement avec leur taille. Les cellules cancéreuses montrent souvent une grande variabilité dans leur forme, contrairement aux cellules normales qui présentent une morphologie plus uniforme. Cette variabilité morphologique est le résultat de la croissance anarchique des cellules cancéreuses et des altérations génétiques qui perturbent leur structure et leur organisation. En effet, une grande variation dans la forme des cellules, combinée à une variabilité de taille, est souvent associée à des tumeurs malignes, alors qu'une morphologie plus uniforme est généralement observée dans les tissus normaux ou bénins.[ix]

2.2.4 Marginal_adhesion

Dans le cas de l'adhésion marginale, les cellules normales présentent une tendance naturelle à se maintenir ensemble, formant des structures tissulaires cohésives. Cette adhésion cellulaire est cruciale pour le maintien de l'intégrité et de la fonctionnalité des tissus dans l'organisme. Les cellules normales communiquent entre elles à travers des liaisons cellulaires spécialisées, telles que les jonctions serrées et les jonctions adhérentes, qui les maintiennent étroitement attachées les unes aux autres. En revanche, dans le cas des cellules cancéreuses, cette capacité d'adhésion est souvent altérée. Les mutations génétiques et les modifications épigénétiques associées au processus de carcinogenèse peuvent perturber les mécanismes moléculaires responsables de l'adhésion cellulaire, entraînant une perte de cohésion entre les cellules. Ainsi, les cellules cancéreuses ont tendance à se détacher les unes des autres et à se disperser de manière désordonnée dans les tissus environnants ou à migrer vers d'autres sites du corps. Cette perte d'adhésion cellulaire est un signe caractéristique de malignité.[x]

2.2.5 Epithelial_size

La taille des cellules épithéliales est étroitement liée à l'uniformité de la forme et de la taille cellulaire évoquée précédemment. Les cellules épithéliales normales présentent généralement une taille uniforme et une morphologie cohérente dans un tissu donné. Cependant, dans le cas des cellules épithéliales malignes, des changements significatifs de taille peuvent se produire. Les cellules cancéreuses peuvent devenir notablement agrandies en raison de leur croissance rapide et de leur capacité à échapper aux mécanismes de régulation de la taille cellulaire normale.[xi]

2.2.6 Bare_nucleoli

Le terme "noyaux nus" est utilisé pour décrire une observation cytologique où les noyaux cellulaires sont visibles sans être entourés par le cytoplasme, la substance cellulaire qui remplit l'espace entre le noyau et la membrane cellulaire. Cette caractéristique est souvent associée aux cellules observées dans les tumeurs bénignes. Dans les tissus normaux et dans de nombreux types de tumeurs bénignes, les noyaux cellulaires sont généralement entourés de cytoplasme, donnant aux cellules une apparence homogène

et délimitée. Cependant, dans certaines circonstances, notamment dans les tumeurs bénignes, les noyaux peuvent apparaître plus grands et plus proéminents, et il peut sembler qu'ils occupent une grande partie de l'espace cellulaire, laissant peu ou pas de cytoplasme visible autour d'eux. Cette observation est souvent interprétée comme des "noyaux nus".[xii]

2.2.7 Bland_chromatin

Dans les cellules cancéreuses, la chromatine tend à être plus grossière. La chromatine, qui est constituée d'ADN, d'ARN et de protéines, est la substance à l'intérieur du noyau cellulaire qui contient les gènes et régule leur expression. Dans les cellules normales, la chromatine est souvent décrite comme étant "douce" lorsqu'elle présente une texture uniforme et fine sous le microscope. Cette apparence régulière est généralement observée dans les cellules bénignes, où l'organisation de la chromatine est contrôlée et ordonnée.

En revanche, dans les cellules cancéreuses, des altérations génétiques et épigénétiques peuvent entraîner des changements dans la structure et l'organisation de la chromatine. Ces modifications peuvent conduire à une chromatine plus "grossière" ou plus hétérogène, caractérisée par des zones de condensation de l'ADN et des régions moins organisées. Cette chromatine plus grossière est souvent associée à une activité génétique altérée, y compris une expression anormale des gènes, ce qui peut contribuer à la progression du cancer.[xiii]

2.2.8 Normal_nucleoli

Les nucléoles sont des structures intranucléaires importantes pour la synthèse des ribosomes et la régulation de l'activité cellulaire. Dans les cellules normales, les nucléoles sont généralement de petite taille et peuvent ne pas être visibles à moins d'une observation minutieuse au microscope. Cependant, dans les cellules cancéreuses, les nucléoles peuvent devenir plus proéminents et plus nombreux. Cette augmentation de taille et de nombre des nucléoles est souvent associée à une activité cellulaire accrue, caractéristique des processus de prolifération cellulaire incontrôlée observés dans le cancer.

Plus précisément, dans les cellules cancéreuses, la croissance rapide et la prolifération désordonnée entraînent une demande accrue en ribosomes et en ARN ribosomal, nécessaires à la synthèse protéique. Pour répondre à cette demande, les cellules cancéreuses augmentent la taille et le nombre de leurs nucléoles pour accroître leur capacité de synthèse ribosomale. Ces nucléoles agrandis et plus nombreux deviennent alors visibles lors de l'examen cytologique ou histologique des tissus, fournissant un indice diagnostique potentiel de malignité.[xiv]

2.2.9 Mitoses

La mitose est un processus crucial dans le cycle de vie cellulaire, où une cellule mère se divise en deux cellules filles identiques. Ce processus est essentiel pour la croissance, le développement et la réparation des tissus dans l'organisme. Il se déroule en plusieurs étapes distinctes, notamment la prophase, la métaphase, l'anaphase et la télophase, suivies de la cytokinèse, où le cytoplasme est divisé pour former deux cellules filles distinctes.

Dans le contexte du cancer, la mitose revêt une importance particulière car les cellules cancéreuses ont souvent une capacité de division cellulaire accrue par rapport aux cellules normales. Cette prolifération cellulaire incontrôlée est l'une des caractéristiques fondamentales du cancer. Les pathologistes peuvent examiner les échantillons de tissus sous un microscope pour évaluer le nombre de mitoses présentes dans les cellules cancéreuses. Un nombre élevé de mitoses, indiqué par la présence de cellules en division, peut être un indicateur de l'activité proliférative et de l'agressivité du cancer.[xv]

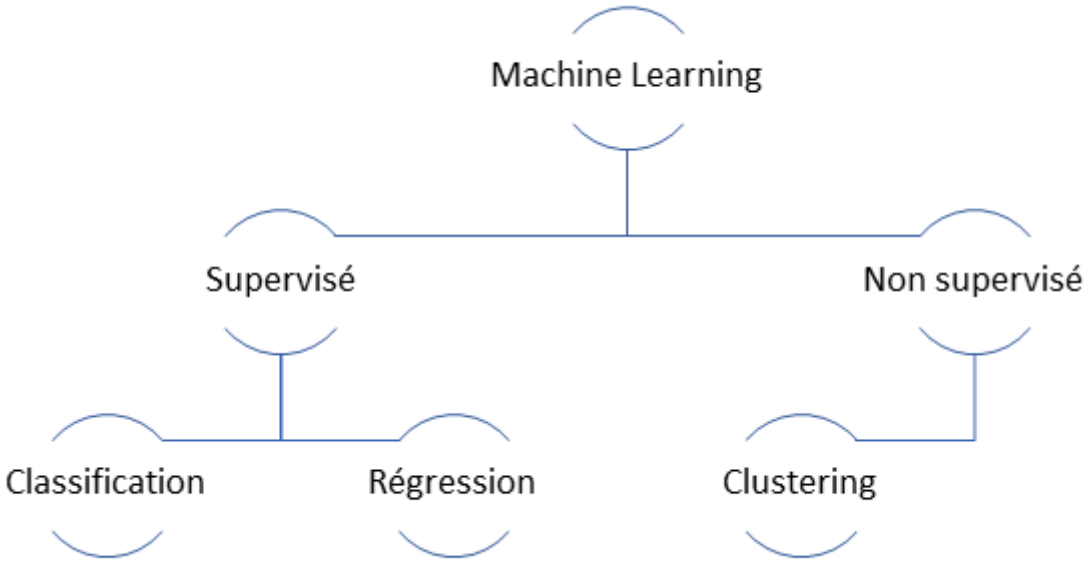
3. Modèles de Classification

3.1 Définition d'un modèle de classification

La classification est un processus d'organisation de données en catégories ou en groupes sur la base de leurs caractéristiques communes.

Dans le domaine du Machine Learning, la classification consiste à entraîner un modèle (appelé aussi algorithme) à reconnaître et à assigner automatiquement des étiquettes ou des classes à de nouvelles données.

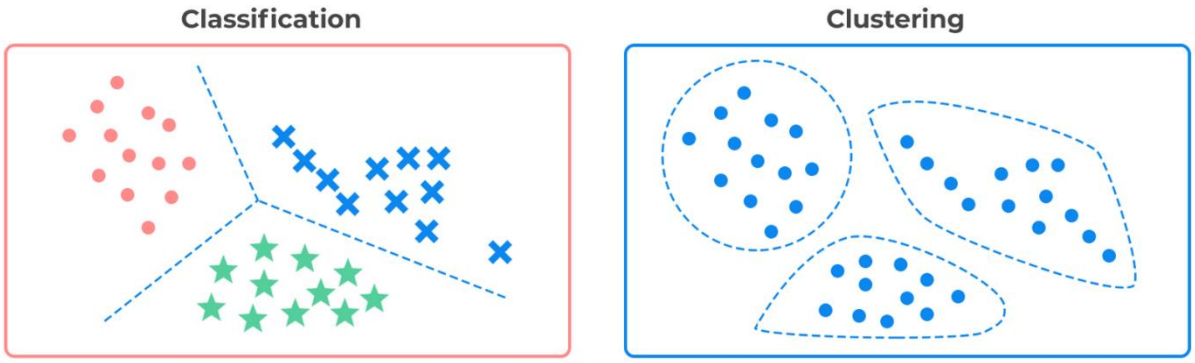
Figure 5 : Segmentation des modèles de Machine Learning



La classification fait partie des méthodes d'apprentissage supervisé, c'est-à-dire que les prédictions que l'on observe en sortie, sont le fruit de l'entrainement d'un modèle sur des données d'entrainement. Ces données d'entraînement utilisées pour former le modèle comprennent à la fois les caractéristiques des données et les étiquettes correspondantes qui indiquent la réponse attendue.

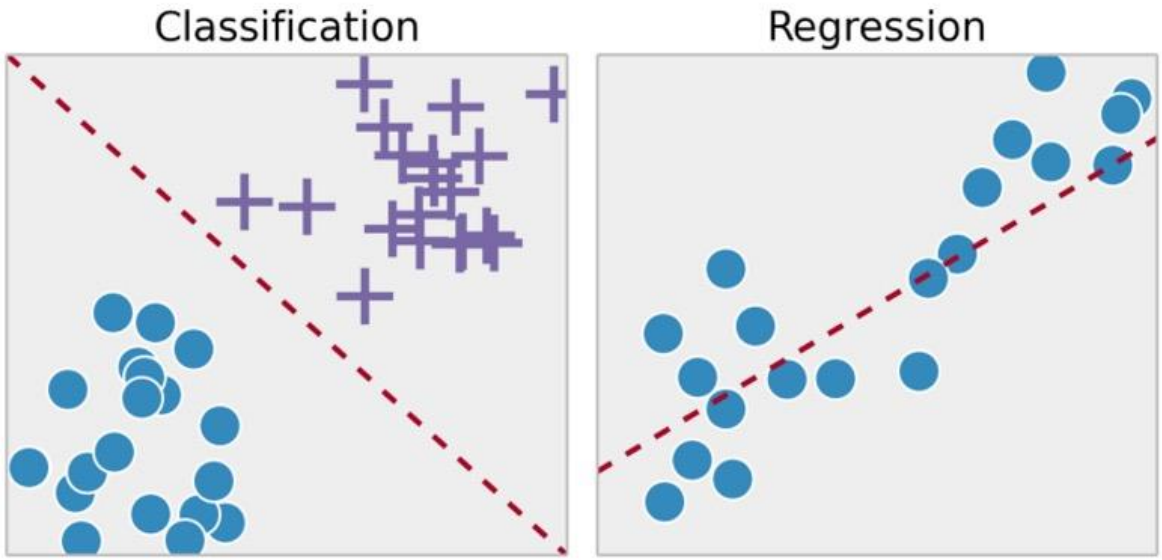
A l'inverse, l'apprentissage non supervisé est utilisé pour tirer des conclusions et trouver des tendances à partir de données d'entrée sans étiquettes comme le clustering.

Figure 6 : Classification versus Clustering



Enfin, la différence essentielle entre les modèles de classification et de régression réside dans la nature des données. Les modèles de classification sont utilisés pour prédire des catégories ou des classes, tandis que les modèles de régression prédisent des valeurs numériques. Les premiers sont adaptés aux variables cibles qualitatives, tandis que les seconds sont utilisés pour les variables cibles quantitatives.

Figure 7 : Classification versus Clustering



3.2 Choisir les modèles adaptés

Dans l'optique de choisir le meilleur modèle pour son jeu de données, il existe plusieurs critères de sélection.

3.2.1 La quantité de données

Le chiffre retenu pour qualifier un jeu de données comme volumineux est en général 100 000 points.

Il existe des modèles adaptés aux jeux de données volumineux comme les SGDclassifier qui utilisent la méthode de descente de gradient[xvi] pour estimer les paramètres du modèle de manière itérative, en utilisant des mini-lots de données à chaque étape de la descente de gradient, ce qui permet un entraînement efficace même sur des ensembles de données massifs.

D'autres modèles seraient inutilisables sur de tels jeux de données, comme par exemple l'algorithme des k plus proches voisins (k-NN) qui nécessite le calcul de la distance entre un point et tous les autres points du jeu de données, ce qui peut être prohibitif en termes de temps et de ressources pour de grands ensembles de données.

3.2.2 La structure des données

En Machine Learning, les données telles que les images, les sons ou les textes sont catégorisées comme "non structurées" et nécessitent souvent l'utilisation de réseaux de neurones en raison de leur efficacité dans la détection de motifs complexes. En revanche, les données tabulaires, plus familières et structurées, peuvent être traitées avec n'importe quel modèle de ML.

3.2.3 La normalité des données

Les modèles paramétriques tels que la régression logistique et les réseaux de neurones sont souvent plus performants lorsque les données présentent une distribution normale, tandis que les modèles non-paramétriques tels que les k-NN et les arbres de décision peuvent s'adapter à une plus grande variété de distributions de données.

Pour remédier à ce problème, il est courant d'utiliser des techniques de prétraitement des données telles que la normalisation ou la standardisation afin de rendre les données plus

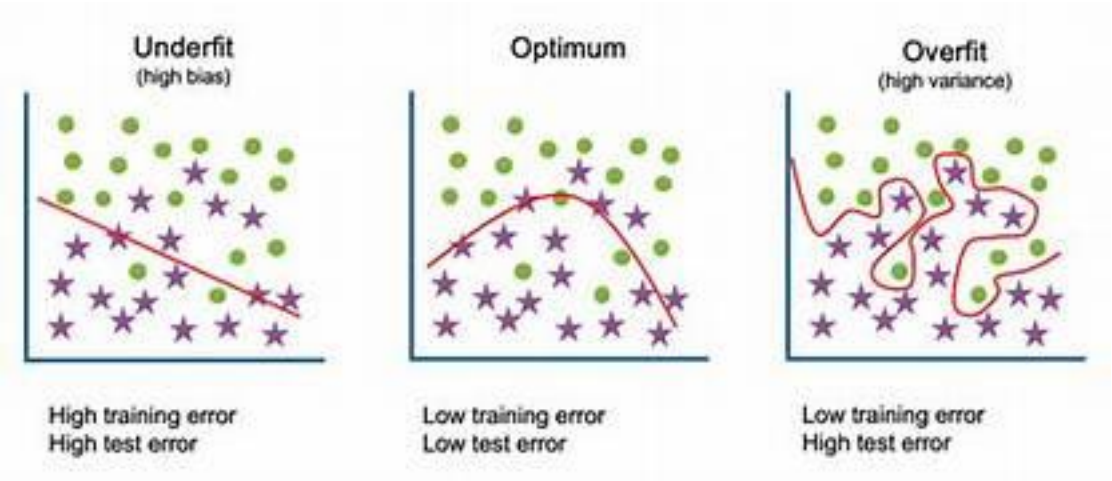
compatibles avec les hypothèses des modèles, en particulier pour les modèles paramétriques qui supposent souvent une distribution normale des données.

3.2.4 La nature des variables

Certains modèles, comme les arbres de décision, montrent une moins grande efficacité lorsque le jeu de données contient un grand nombre de variables quantitatives. Les arbres de décision sont conçus pour diviser les axes de manière orthogonale, ce qui peut entraîner un risque de surajustement (overfitting) élevé lorsqu'ils sont appliqués à des relations linéaires probables avec des variables quantitatives. Le surajustement survient lorsque le modèle s'adapte excessivement aux données d'entraînement, ce qui compromet sa capacité à généraliser sur de nouvelles données et peut conduire à des performances médiocres et des décisions erronées.

En revanche, ces modèles sont généralement performants avec des variables qualitatives.

Figure 8 : Sous-ajustement et Surajustement



3.3 Modèles utilisés pour l'application

3.3.1 Régression logistique

La régression logistique fait partie des modèles linéaires et contrairement à son nom est utilisée pour la classification plutôt que pour la régression.

Le principe de base de la régression logistique est de modéliser la probabilité que les données appartiennent à une classe particulière en utilisant une fonction logistique. Cette fonction logistique transforme les valeurs d'entrée en une plage de valeurs entre 0 et 1, représentant les probabilités. Plus précisément, la régression logistique modélise la relation entre les variables indépendantes (caractéristiques) et une variable dépendante binaire (classe cible) en utilisant une fonction logistique pour estimer les probabilités de chaque classe. Ici, les variables indépendantes représentent les caractéristiques observées sur les cellules tumorales, telles que leur taille, leur forme ou encore leur adhérence. La variable dépendante indique si la tumeur est bénigne ou maligne.

Lors de l'entraînement du modèle, les coefficients sont ajustés de manière à maximiser la probabilité d'observer les étiquettes de classe réelles pour les données d'entraînement. Une fois le modèle entraîné, il peut être utilisé pour prédire la probabilité qu'une nouvelle observation appartienne à chaque classe, et la classe avec la probabilité la plus élevée est souvent choisie comme prédiction finale.[xvii]

La régression logistique est adaptée aux problèmes de classification binaire (deux classes), mais elle peut également être étendue à des problèmes de classification multi-classe en utilisant des techniques telles que la méthode "One-vs-Rest" ou la régression logistique multinomiale. Dans notre cas, nous utiliserons la régression logistique binaire puisque notre variable dépendante comprend deux classes.

A noter que si l'on avait un nombre de variable élevé, une régression pénalisée aurait pu être utilisé pour régulariser le modèle en introduisant une pénalité dans la fonction de coût du modèle pour limiter la magnitude des coefficients des variables explicatives.

Il existe plusieurs types de pénalités comme celle de Lasso qui favorise la parcimonie en encourageant certains coefficients à devenir exactement nuls, ce qui peut conduire à une sélection de variables.[xviii]

3.3.2 Arbres de décision

Les modèles basés sur les arbres de décision sont des algorithmes d'apprentissage supervisé utilisés pour la classification et la régression. Le principe de base est de diviser récursivement l'ensemble de données en sous-ensembles plus petits en fonction des caractéristiques des données. Cela se fait en sélectionnant itérativement les caractéristiques qui divisent le mieux les données en groupes homogènes, c'est-à-dire des groupes où les observations partagent des caractéristiques similaires dans la classe cible. Ces divisions se font en utilisant des seuils sur les caractéristiques, et les décisions de division sont prises de manière à maximiser la pureté des sous-groupes résultants.

Une fois que l'arbre de décision est construit, il peut être utilisé pour faire des prédictions en appliquant les règles de décision apprises à de nouvelles observations. Pour une classification, l'observation est acheminée à travers les nœuds de l'arbre en suivant les règles de décision basées sur les caractéristiques, jusqu'à ce qu'elle atteigne une feuille qui correspond à une classe prédite.

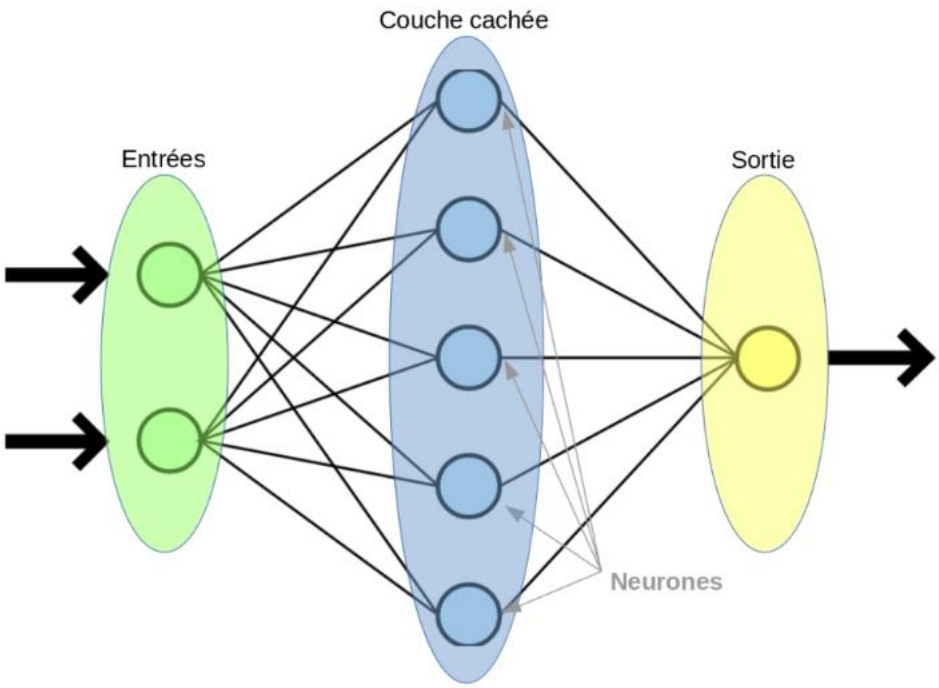
Les arbres de décision sont attrayants en raison de leur interprétabilité, de leur facilité d'utilisation et de leur capacité à gérer des données complexes et hétérogènes. Cependant, ils ont tendance à être sensibles au surapprentissage, ce qui peut entraîner une faible généralisation à de nouvelles données. Pour atténuer ce problème, des techniques telles que la taille maximale de l'arbre, la taille minimale des feuilles, la réduction de la complexité de l'arbre (élagage) sont souvent utilisées.[xix]

3.3.3 Réseaux de neurones

Les réseaux de neurones sont des modèles d'apprentissage automatique inspirés du fonctionnement du cerveau humain, particulièrement puissants pour les tâches complexes comme la reconnaissance d'images, la traduction automatique et les jeux. Leur structure de base comprend des unités de calcul appelées neurones, organisées en couches : une couche d'entrée pour les données brutes, des couches cachées pour les transformations intermédiaires, et une couche de sortie pour produire les résultats finaux.

Il existe plusieurs types de réseau de neurones utiles pour des tâches diverses et variées. Nous utiliserons le Perceptron Multicouche (MLP) qui est un type de réseau de neurones simple et puissant, utilisé principalement pour des tâches de classification et de régression. Il est composé de plusieurs couches de neurones, incluant une couche d'entrée pour recevoir les données (caractéristiques cellulaires en vert), une ou plusieurs couches cachées pour effectuer des transformations intermédiaires, et une couche de sortie pour produire les résultats finaux (tumeur bénigne ou maligne en jaune).

Figure 9 : Schéma d'un réseau de neurones



Chaque neurone dans une couche est connecté à tous les neurones de la couche suivante, ce qui permet au MLP de capturer des relations non linéaires entre les variables d'entrée et de sortie. L'entraînement du MLP se fait par propagation avant des données à travers le réseau, évaluation de la perte, et rétropropagation pour ajuster les poids des connexions afin de minimiser l'erreur.

Bien que le MLP soit adapté pour des tâches simples, sa structure entièrement connectée peut rendre l'entraînement inefficace pour des données très volumineuses ou complexes. Etant donné que nos données ne posent pas ce problème, son utilisation semble appropriée.

Les avantages des réseaux de neurones incluent leur capacité à capturer des relations complexes, leur généralisation à de nouvelles données, et leur adaptabilité à diverses tâches. Cependant, ils nécessitent temps de calcul importants et sont souvent considérés comme des "boîtes noires"[4], ce qui complique l'interprétation des décisions.[xx]

[4] Le terme "boîte noire" est utilisé pour décrire des modèles ou des systèmes dont le fonctionnement interne n'est pas facilement compréhensible ou interprétable par les humains.

4. Application empirique

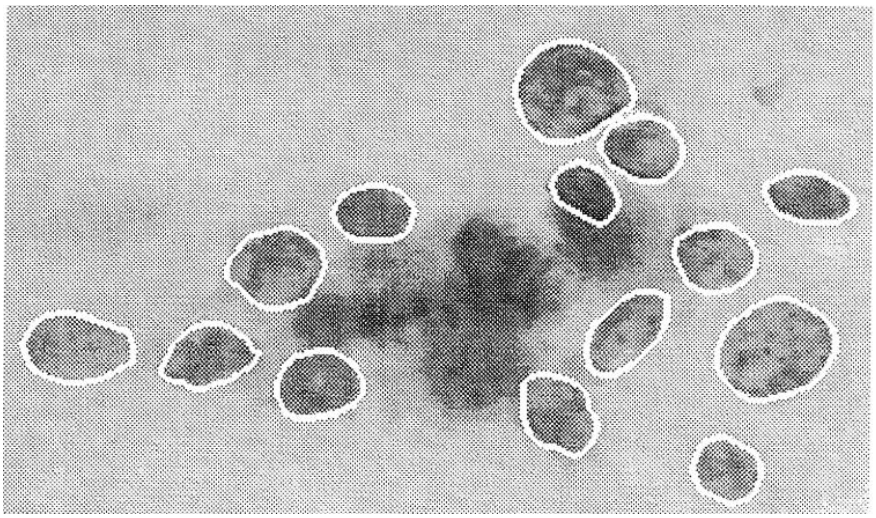
Il est important de souligner que tous les codes et figures présentés dans cette section ont été créés à l'aide du logiciel RStudio, sauf indication contraire.

4.1 Présentation des données

Le jeu de données utilisé dans cette partie est disponible publiquement[xxi] et a été créé par le Dr William H. Wolberg, médecin à l'hôpital de l'Université du Wisconsin à Madison, aux États-Unis. Il a été donné par Olvi Mangasarian le 15 juillet 1992. Pour créer le jeu de données, le Dr Wolberg a utilisé des échantillons de liquide prélevés sur des patients présentant des masses mammaires solides et un programme informatique graphique développé quelques années plus tôt appelé « snakes »[xxii], capable d'effectuer l'analyse des caractéristiques cytologiques basée sur de la numérisation.

Les "snakes" sont des outils utilisés en traitement d'images pour effectuer des contours ou des délimitations précises autour des objets dans une image. Ils ont été développés pour suivre des contours en imagerie médicale et dans d'autres applications de vision par ordinateur. Cet outil fonctionne en initiant un contour initial (généralement proche du contour de l'objet d'intérêt) et en le déformant de manière itérative pour s'adapter à la structure de l'image.[xxiii]

Figure 10 : Représentation des contours des noyaux cellulaires utilisant les « snakes »[5]



[5] Image issue de l'étude citée précédemment : Nuclear feature extraction for breast tumor diagnosis. Electronic imaging.

Le programme utilise un algorithme de courbe d'ajustement, tel qu'illustré dans la Figure 10, pour calculer dix caractéristiques de chaque cellule de l'échantillon. Ensuite, il détermine la moyenne, la valeur extrême et l'erreur standard de chaque caractéristique pour l'image, produisant ainsi un vecteur de 30 valeurs réelles.

Le système compare également différentes caractéristiques pour chaque noyau. Chaque caractéristique est évaluée sur une échelle de 1 à 10, avec 1 représentant le plus proche du bénin et 10 le plus proche du malin. Une analyse statistique a identifié neuf caractéristiques significativement différentes entre les échantillons bénins et malins : uniformité de la forme et de la taille des cellules, épaisseur des amas, présence de noyaux nus, taille cellulaire, présence de nucléoles normaux, cohésion des amas, chromatine nucléaire et mitoses.[6] La table 1 résume l'état actuel de l'ensemble de données utilisé dans cet article.

Table 1 : Résumé du jeu de données

Variables	**Type**	**Répartition**
Clump_thickness	Numérique	1-10
Size_uniformity	Numérique	1-10
Shape_uniformity	Numérique	1-10
Marginal_adhesion	Numérique	1-10
Epithelial_size	Numérique	1-10
Bare_nucleoli	Numérique	1-10
Bland_chromatin	Numérique	1-10
Normal_nucleoli	Numérique	1-10
Mitoses	Numérique	1-10
Class	Nominale	Bénigne-Maligne
Nombre d'observations initiales	701	

Les échantillons ont été prélevés périodiquement lors du signalement par le Dr Wolberg de ses cas cliniques, ce qui signifie que les données sont regroupées chronologiquement

[6] Le nom des variables est repris en anglais dans la table et les codes utilisés sur R. La traduction de chaque variable et des détails supplémentaires sur chacune d'elles sont disponibles dans la partie 2.2.

selon leur date de création. La Table illustre le nombre d'instances[7] ajoutées chaque mois depuis la création de l'ensemble de données (janvier 1989) jusqu'à la dernière instance créée (novembre 1991).

Avant sa publication publique, l'ensemble de données comptait 701 observations. Cependant, en janvier 1989, après une révision, deux instances du groupe 1 ont été jugées incohérentes et ont été retirées.

Table 2 : Groupes des instances

Groupe 1	369	Janvier 1989
Groupe 2	70	Octobre 1989
Groupe 3	31	Février 1990
Groupe 4	17	Avril 1990
Groupe 5	48	Août 1990
Groupe 6	49	Janvier 1991
Groupe 7	31	Juin 1991
Groupe 8	86	Novembre 1991
Total : 701		

[7] En médecine, une « instance » correspond à un cas clinique signalé, c'est-à-dire à une observation individuelle.

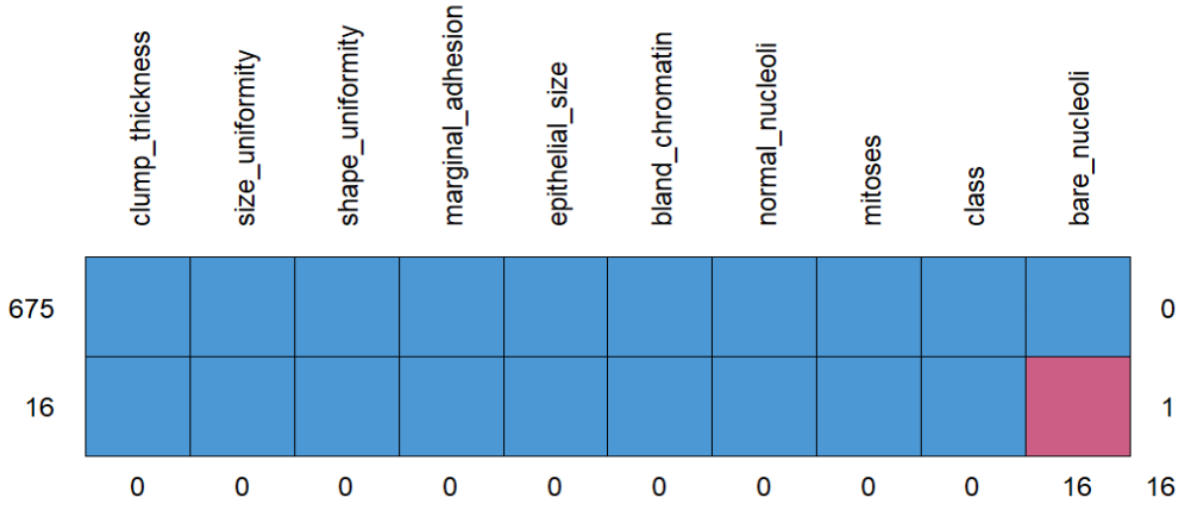
4.2 Prétraitement des données

Tout d'abord, l'identification et la gestion des doublons sont cruciales pour éviter des résultats biaisés et garantir l'exactitude des analyses. Les valeurs manquantes, quant à elles, peuvent réduire la puissance statistique et mener à des conclusions erronées si elles ne sont pas traitées adéquatement. Gérer ces deux aspects améliore la fiabilité des modèles et la qualité des décisions basées sur les données.

Table 3 : Doublons et Valeurs manquantes

Doublons	Valeurs manquantes
8	16

Figure 11 : Valeurs manquantes

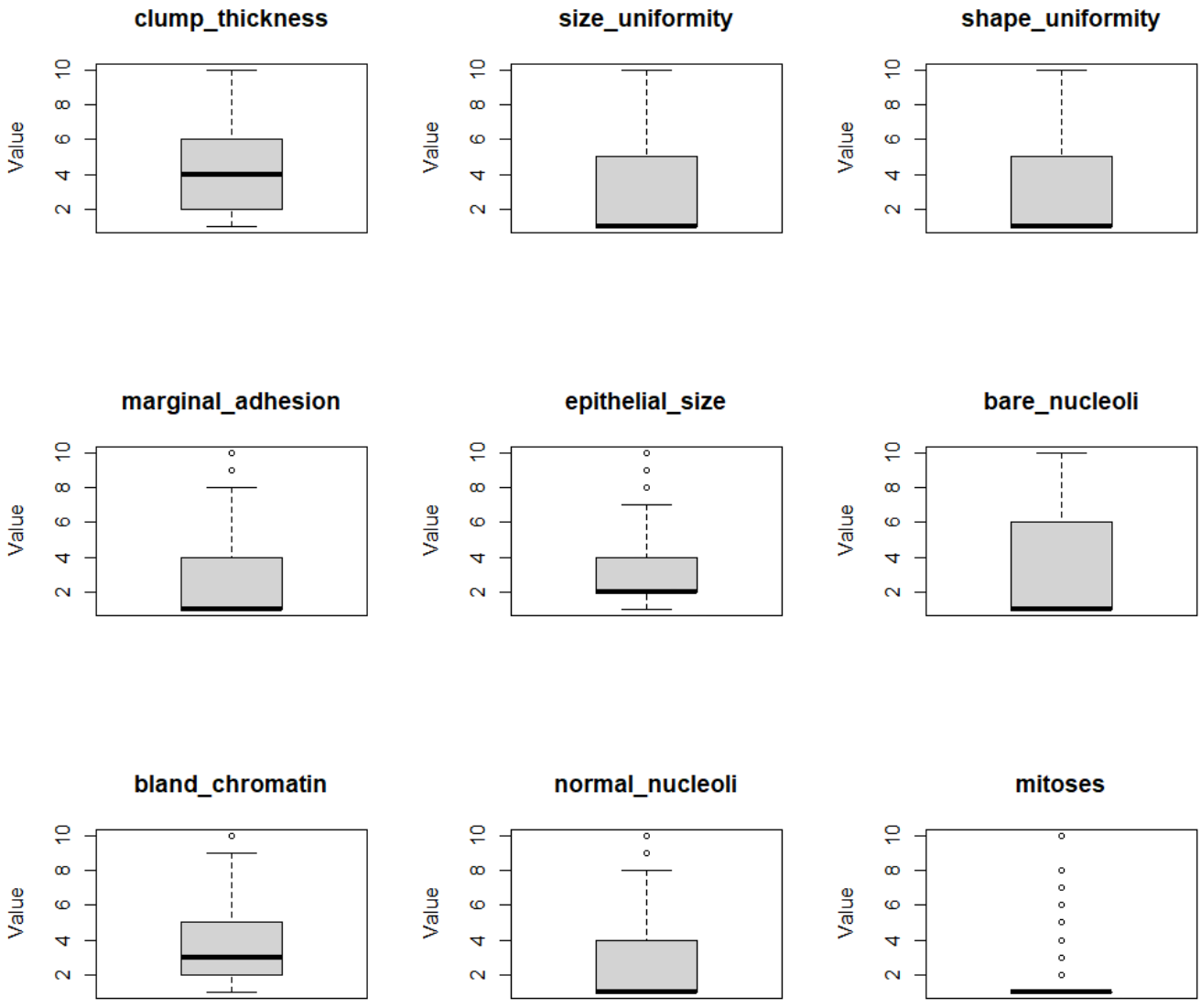


Notre base de données comporte huit doublons qui seront supprimés. En ce qui concerne les valeurs manquantes, elles se trouvent toutes dans la variable bare_nuleoli. Plusieurs options s'offrent à nous pour les traiter. Nous pouvons, par exemple, utiliser des méthodes d'imputation telles que le "predictive mean matching", qui remplace les valeurs manquantes par des valeurs prédites en fonction des moyennes des données similaires. Alternativement, nous pouvons choisir de supprimer les observations contenant des valeurs manquantes et c'est précisément cette approche qui a été retenue. Bien que cela réduise la taille de notre échantillon et puisse potentiellement affecter la représentativité des résultats, cela garantit la fiabilité et la précision des

analyses. En effet, l'imputation des valeurs manquantes peut introduire des biais indésirables, ce qui est particulièrement problématique dans le domaine de la santé, où de telles erreurs peuvent avoir des conséquences importantes.

Ensuite, les valeurs aberrantes, également appelées valeurs extrêmes, sont des observations qui diffèrent considérablement du reste de l'ensemble de données. Elles peuvent résulter d'erreurs de mesure, de saisie incorrecte des données ou de phénomènes rares mais légitimes. Les valeurs aberrantes peuvent fausser les analyses statistiques et conduire à des conclusions erronées si elles ne sont pas correctement identifiées et traitées.

Figure 12 : Boîtes à moustaches des variables explicatives



D'un point de vue strictement statistique, l'application du test de Rosner sur chaque variable soupçonnée de contenir des valeurs aberrantes a révélé que seules les valeurs de la variable "Mitoses" étaient identifiées comme des outliers (annexe).

Cependant, les valeurs aberrantes dans les données de santé peuvent parfois fournir des informations précieuses et ne reflètent pas nécessairement des erreurs ou des anomalies. Elles peuvent représenter des cas extrêmes ou rares ayant une importance clinique. Supprimer les valeurs aberrantes pourrait potentiellement entraîner la perte d'informations précieuses et affecter l'exactitude et la fiabilité de l'analyse.

Elles peuvent également fournir des informations précieuses à des fins de diagnostic, notamment dans les situations où elles représentent des maladies rares ou des symptômes inhabituels. Supprimer les valeurs aberrantes pourrait entraver la capacité à détecter et diagnostiquer avec précision de tels cas, c'est pourquoi nous choisissons de conserver toutes les observations pour la suite de l’analyse.

4.3 Analyse statistique et descriptive

La section suivante se concentre sur l'analyse statistique descriptive de notre ensemble de données. À travers cette analyse, nous chercherons à obtenir un aperçu approfondi des caractéristiques de nos variables ainsi que de la distribution de nos données. Cette exploration nous permettra de mieux comprendre la nature et la structure de notre jeu de données, posant ainsi les bases pour des analyses plus avancées et des interprétations plus fines.

Table 4 : Répartition des classes de tumeurs

Variables	Effectif
Classe	
Bénigne	65.03 %
Maligne	34.97%

La table 4 indique la fréquence de répartition entre les deux classes de tumeurs et l'on observe une majorité de tumeurs bénignes : 439 contre 236 malignes.

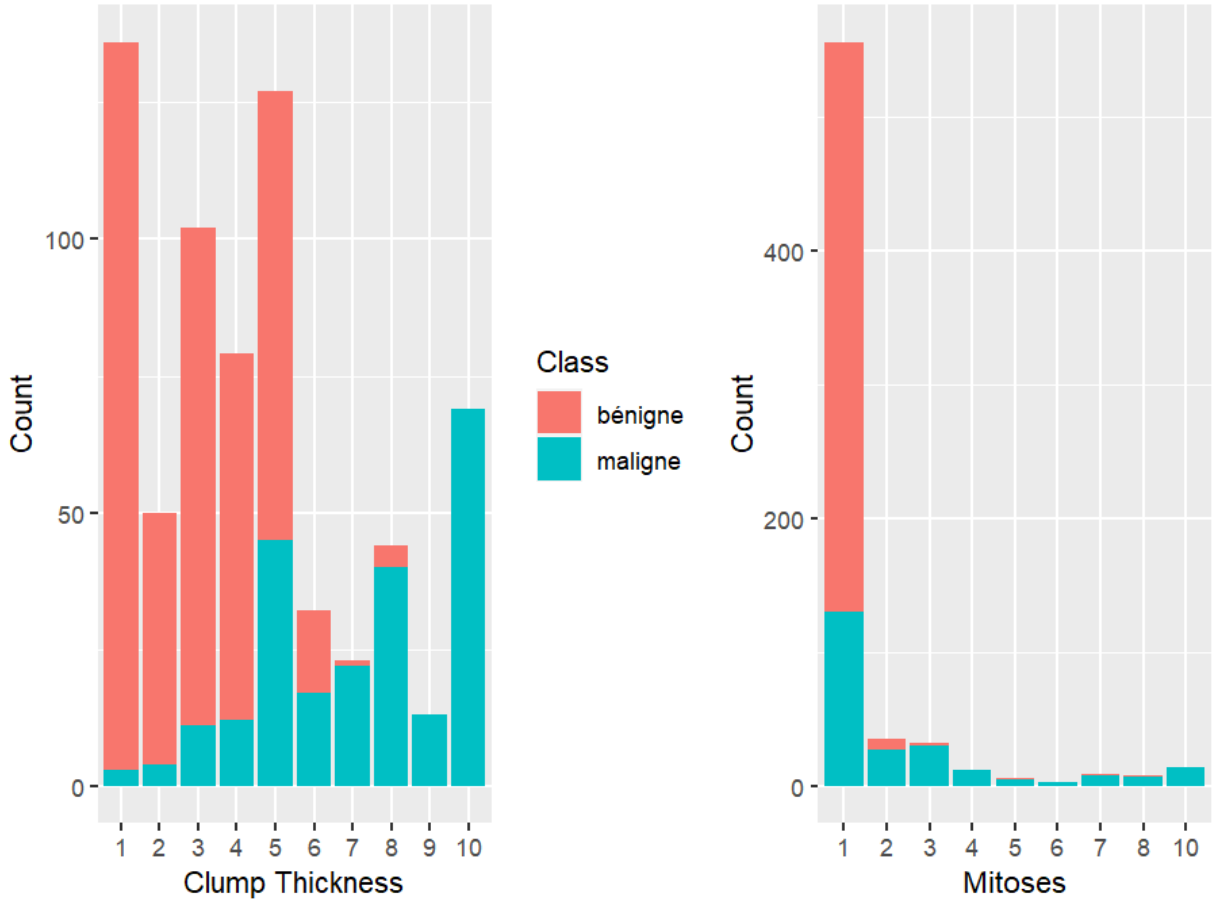
Table 5 : Statistiques descriptives des variables explicatives

Variables	Min	1st qu	Median	Mean	3st Qu	Max	Ecart-type
Clump_thickness	1	2	4	4.452	6	10	2.82
Size_uniformity	1	1	1	3.147	5	10	3.05
Shape_uniformity	1	1	1	3.209	5	10	2.97
Marginal_adhesion	1	1	1	2.849	4	10	2.87
Epithelial_size	1	2	2	3.23	4	10	2.20
Bare_nucleoli	1	1	1	3.538	6	10	3.63
Bland_chromatin	1	2	3	3.443	5	10	2.45
Normal_nucleoli	1	1	1	2.886	4	10	3.06
Mitoses	1	1	1	1.607	1	10	1.74

Nous observons que les statistiques des variables sont généralement basses, avec des moyennes inférieures à 5 sur une échelle de 1 à 10. Cette tendance est cohérente avec la répartition des classes précédemment examinée, qui montre une prédominance des tumeurs bénignes. En effet, sur l'échelle de 1 à 10 des variables, une valeur plus faible correspond à une caractéristique moins maligne, ce qui est en accord avec la prévalence des tumeurs bénignes dans notre ensemble de données.

Pour observer plus clairement la répartition des données, nous pouvons réaliser des diagrammes en barres des caractéristiques cellulaires selon leur classe.

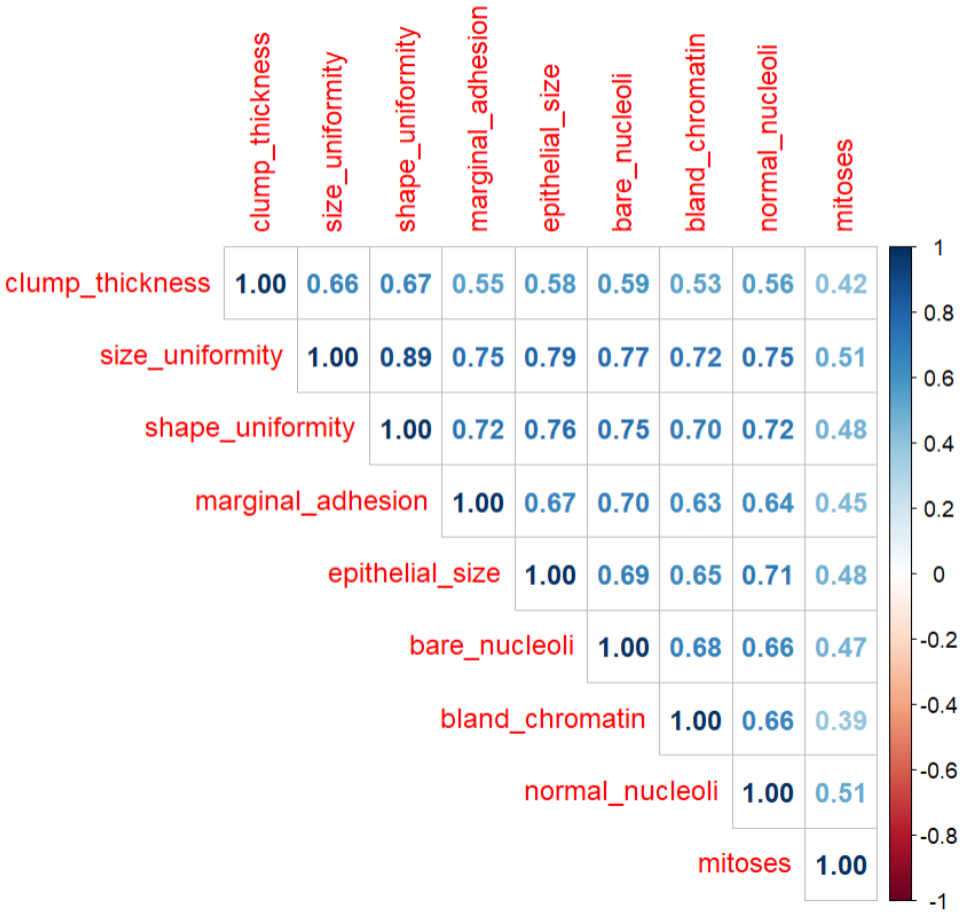
Figure 13 : Diagrammes en barres de l'épaisseur de l'amas et des mitoses selon leur classe



L'épaisseur des amas (Clump Thickness) reflète la majorité des répartitions des classes de valeurs, c'est pourquoi nous avons choisi de ne pas afficher les autres variables. Quant aux mitoses, elles présentent une distribution particulièrement singulière dans ce jeu de données, avec les trois premiers quartiles regroupés dans la première catégorie. Cette

distribution très concentrée peut poser des problèmes, car les modèles prédictifs auront potentiellement du mal à capturer la diversité des schémas présents.

Figure 14 : Corrélations des variables explicatives



Toutes les variables présentent des corrélations positives fortes entre elles, indiquant un niveau d'interdépendance élevé entre elles. Cette multicolinéarité peut compliquer l'estimation précise des coefficients de régression, rendant ces coefficients instables et sensibles aux petites modifications des données. En conséquence, il devient difficile d'interpréter les relations entre chaque caractéristique cellulaire et leur classe, car l'effet individuel de chaque caractéristique est obscurci. Cependant, dans le cadre de modèles prédictifs visant uniquement à déterminer la classe de la tumeur, ce problème est moins préoccupant car nous ne nous intéressons pas à l'interprétation des coefficients des variables.

4.4 Analyse des résultats

Dans cette partie, trois modèles prédictifs sont présentés : un modèle de régression logistique, un modèle utilisant des arbres de décisions ainsi qu'un modèle de réseaux de neurones.

Nous sommes en présence de modèle simples, et pour certains paramétriques. Pour des raisons de performance, il a été décidé de normaliser les données (ajustement des valeurs des caractéristiques à une échelle commune) par la méthode min-max (mise à l'échelle entre 0 et 1).

Nous avons parlé précédemment de la répartition inégale des classes dans le jeu de données : il y a plus de tumeurs bénignes que malignes. Cela peut poser des problèmes lors de la réalisation de modèles d'apprentissage automatique qui peuvent devenir biaisés envers la classe majoritaire, dans ce cas les tumeurs bénignes. Les algorithmes peuvent avoir tendance à prédire majoritairement cette classe sans tenir compte des caractéristiques réelles des données, ce qui peut conduire à des performances médiocres. Afin de remédier à ce problème, une technique de suréchantillonnage a été utilisé sur la classe minoritaire. A noter que seul l'ensemble d'entraînement est modifié par le suréchantillonnage, tandis que l'ensemble de test reste inchangé. Cela est important pour assurer que l'évaluation des performances du modèle se fait sur un ensemble de test représentatif de la distribution initiale des données.[8]

4.4.1 Régression logistique

Dans notre première tentative de modélisation avec une régression logistique de forme binaire, nous avons employé la bibliothèque glmnet.

[8] Un ensemble d'apprentissage composé de 80% des données initiales a été utilisé. C'est uniquement sur ces 80% que l'on suréchantillonne. Les 20% utilisés pour tester les modèles n'ont pas été modifié.

Modèle RL :

$$Y = \alpha + \beta 1 Clump\_thickness + \beta 2 Size\_uniformity + \beta 3 Shape\_uniformity + \beta 4 Marginal\_adhesion + \beta 5 Epithelial\_size + \beta 6 Bare\_nucleoli + \beta 7 Bland\_chromatin + \beta 8 Normal\_nucleoli + \beta 9 Mitoses + \epsilon$$

Le modèle étant ainsi spécifié, nous supposons pour cette partie de notre travail que toutes les variables explicatives utilisées pour la prédiction de classe sont toutes exogènes, satisfaisant la condition suivante :

$$cov(xi,ei) = 0$$

Nous supposons donc une détermination parfaite du modèle sans erreur de mesure, sans variables manquantes corrélées à la variable expliquée et aux variables explicatives, et enfin, sans équations simultanées.

Table 6 : Modèle de régression logisique

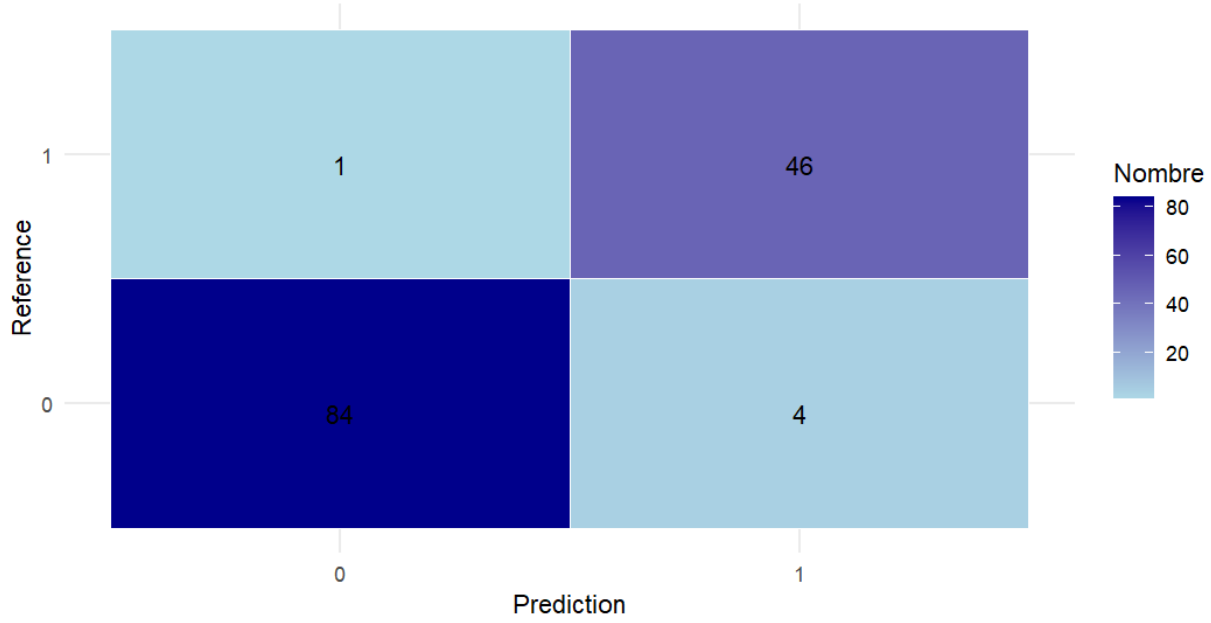
Variables	Coefficient	Pr(>|z|)
Constante	-6.8372	< 2e-16 ***
Clump_thickness	3.8478	0.00127 **
Size_uniformity	2.2982	0.38982
Shape_uniformity	5.5682	0.03271 *
Marginal_adhesion	2.1052	0.07264 .
Epithelial_size	1.4540	0.33231
Bare_nucleoli	4.4766	7.11e-06 ***
Bland_chromatin	1.3054	0.47820
Normal_nucleoli	2.3882	0.04044 *
Mitoses	5.5926	0.05215 .

D'un point de vue économétrique, ce modèle est plutôt bon puisque 4 des 9 caractéristiques cellulaire sont significatives au seuil de 5% (*) et donc en mesure

d’expliquer la classe des tumeurs des individus. Par ordre de coefficient, ces 4 variables sont : la forme de cellule, la présence de noyaux nus, l’épaisseur des amas cellulaire et la présence de nucléoles. De manière générale, une augmentation de ces variables sur l’échelle de 1 à 10 est associée à une augmentation de la probabilité que la tumeur soit classée comme maligne.

Voici désormais une présentation détaillée des résultats obtenus à l'aide de ce modèle de prédiction.

Figure 15 : Matrice de confusion du jeu test, modèle RL



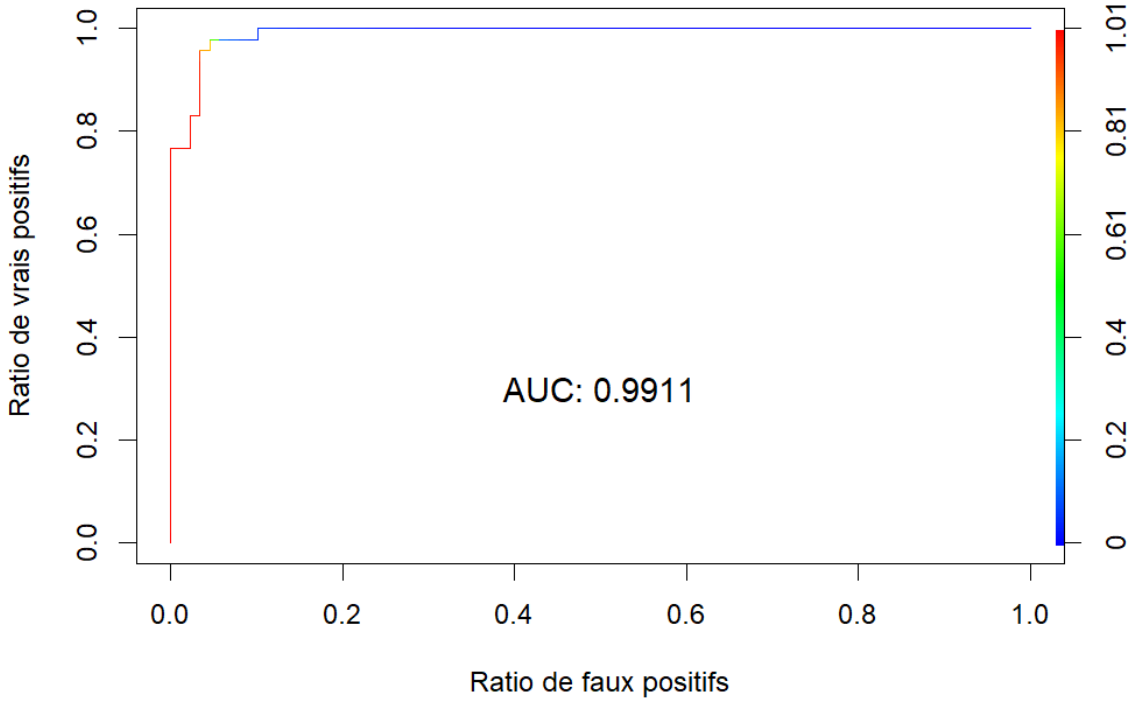
Aussi connue sous le nom de rappel ou taux de vrais positifs, la sensibilité mesure la capacité du modèle à identifier correctement les vrais positifs parmi tous les cas positifs réels. Dans ce cas, la sensibilité est de 0.9787, ce qui signifie que le modèle a identifié correctement 97.87% des tumeurs malignes (classe 1) parmi toutes les tumeurs malignes réelles.

La spécificité mesure quant à elle la capacité du modèle à identifier correctement les vrais négatifs parmi tous les cas négatifs réels. Ici, la spécificité est de 0.9545, ce qui indique que le modèle a identifié correctement 95.45% des tumeurs bénignes (classe 0) parmi toutes les tumeurs bénignes réelles.

Si l'on évalue la performance globale du modèle, la précision de ce dernier est de 0.963, ce qui signifie que le modèle a correctement prédit la classe de 96.3% de toutes les observations dans l'ensemble de test. Cela inclut à la fois les vrais positifs et les vrais négatifs, ainsi que les faux positifs et les faux négatifs.

Nous pouvons observer plus en détail ces résultats grâce à la Courbe ROC (Receiver Operating Characteristic) qui illustre le taux de vrais positifs par rapport au taux de faux positifs entre les deux types de classification.

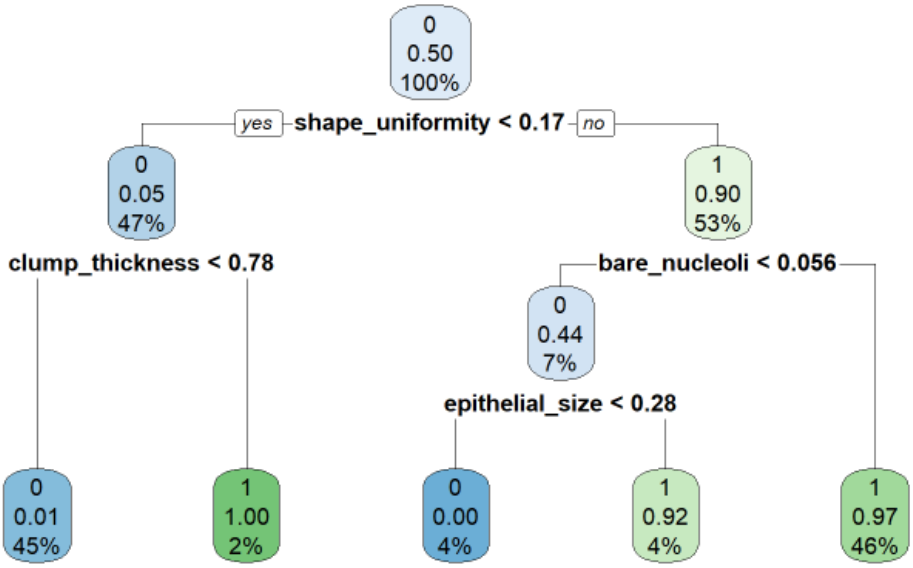
Figure 16 : Courbe ROC, modèle RL



4.4.2 Arbre de décision

Pour la modélisation d'un arbre de décision, une approche largement utilisée est celle offerte par la bibliothèque rpart. Cette bibliothèque propose des outils puissants pour la construction d'arbres de décision à partir de données d'entraînement, permettant ainsi de capturer les relations complexes entre les caractéristiques et la variable cible.

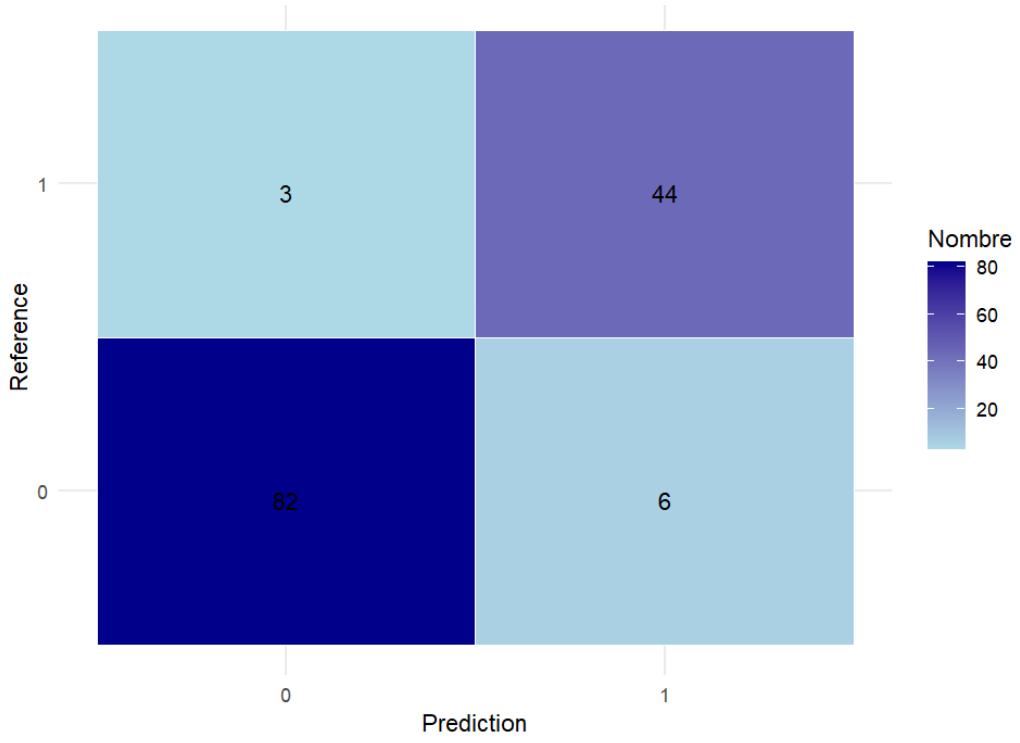
Figure 17 : Arbre de décision



Le nœud racine (qui contient toutes les observations) de notre arbre de décision est basé sur la forme des cellules. L'échantillon se divise donc dans un premier temps en deux groupes, avec à gauche les individus dont la forme des cellules est inférieure à 0.17 sur l'échelle de 0 à 1 (données normalisées), et inversement à droite. Sans rentrer dans les détails, nous pouvons observer que les caractéristiques cellulaires qui composent les nœuds enfants sont l'épaisseur des amas cellulaire, la présence de noyaux nus, et la taille des cellules épithéliales.

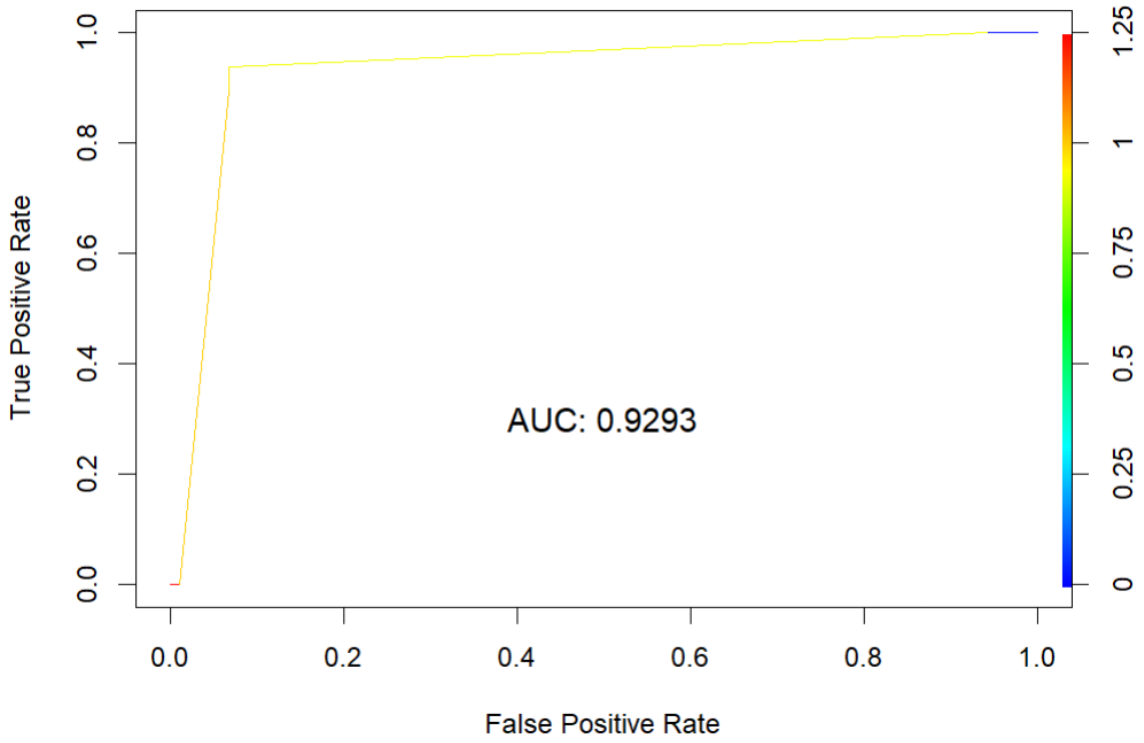
Voici l'analyse détaillée des résultats obtenus avec ce modèle prédictif.

Figure 18 : Matrice de confusion du jeu test, modèle AD



La sensibilité, avec un score de 0.9362, reflète la capacité du modèle à identifier 93.62% des tumeurs malignes parmi toutes celles qui sont réellement malignes. En parallèle, la spécificité, notée à 0.9318, indique que le modèle a correctement identifié 93.18% des tumeurs bénignes parmi toutes celles qui sont réellement bénignes. Globalement, la précision du modèle s'élève à 0.9333, démontrant qu'il a prédit correctement la classe de 93.33% de toutes les observations dans l'ensemble de test.

Figure 19 : Courbe ROC, modèle AD

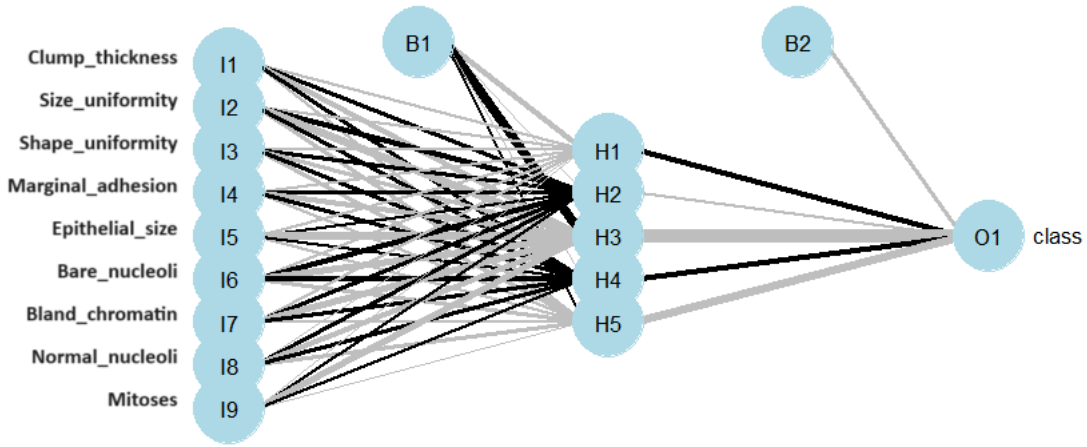


Les performances globales du modèle semblent indiquer une capacité satisfaisante à prédire avec précision les classes des observations dans l'ensemble de test. Cela suggère une certaine fiabilité dans la capacité de ce modèle pour ce qui est généraliser ses prédictions.

4.4.3 Réseaux de neurones

Pour modéliser un réseau de neurones, une approche courante est d'utiliser la bibliothèque nnet. Cette bibliothèque offre des fonctionnalités robustes pour la construction de réseaux de neurones artificiels à partir de données d'entraînement, permettant ainsi de capturer des relations non linéaires entre les variables d'entrée et de sortie. En utilisant nnet, il est possible de spécifier la structure du réseau, y compris le nombre de couches cachées et le nombre de neurones par couche, offrant ainsi une flexibilité pour modéliser des tâches complexes dans divers domaines d'application.

Figure 20 : Réseau de neurones



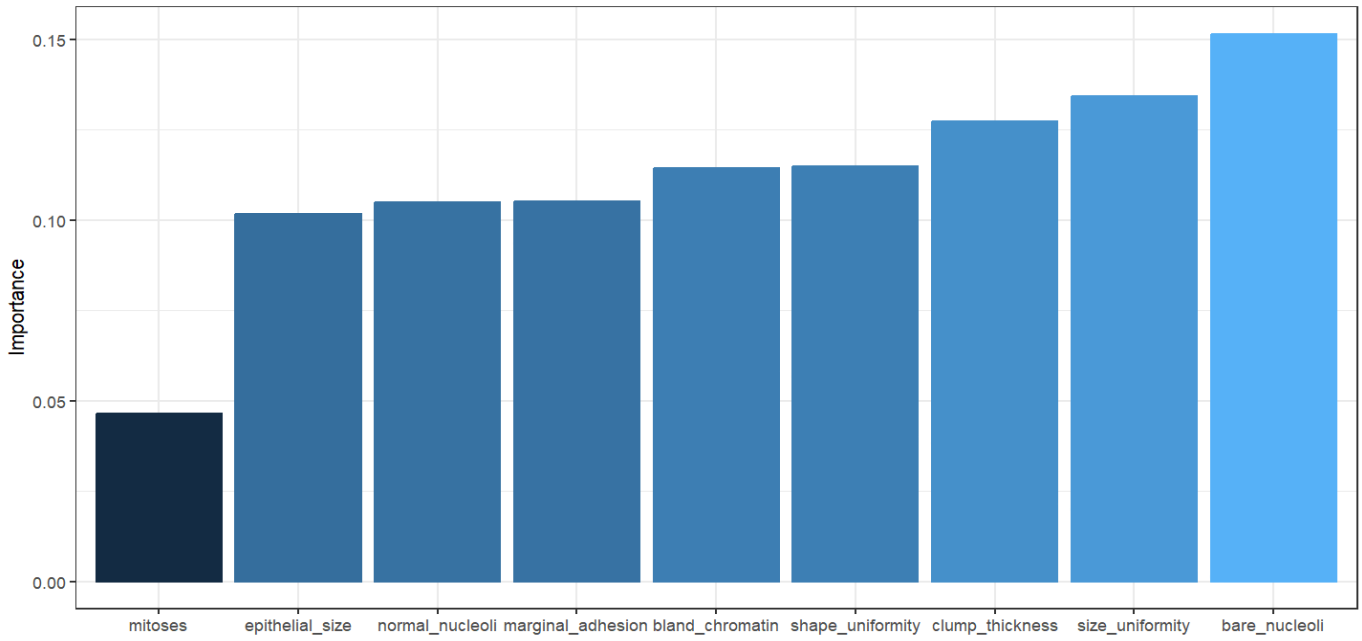
Chaque lettre H suivie d'un numéro représente une couche cachée spécifique (H pour hidden). Chaque couche cachée est composée de plusieurs neurones, et les connexions entre les neurones des couches cachées sont représentées par des flèches qui indiquent le flux de données à travers le réseau. Les couches cachées sont responsables de l'extraction et de la transformation des caractéristiques des données d'entrée, ce qui permet au réseau de neurones de capturer des relations.

Les lettres B1 et B2 représentent les biais (B pour bias) associés à chaque couche cachée du réseau. Les biais sont des valeurs ajoutées à la somme pondérée des entrées avant

l'application de la fonction d'activation dans chaque neurone d'une couche cachée. Ils permettent au réseau de neurones de mieux s'adapter aux données.

Maintenant que nous avons exploré la construction et le fonctionnement de notre modèle de réseau de neurones, passons à l'examen des résultats de nos prédictions

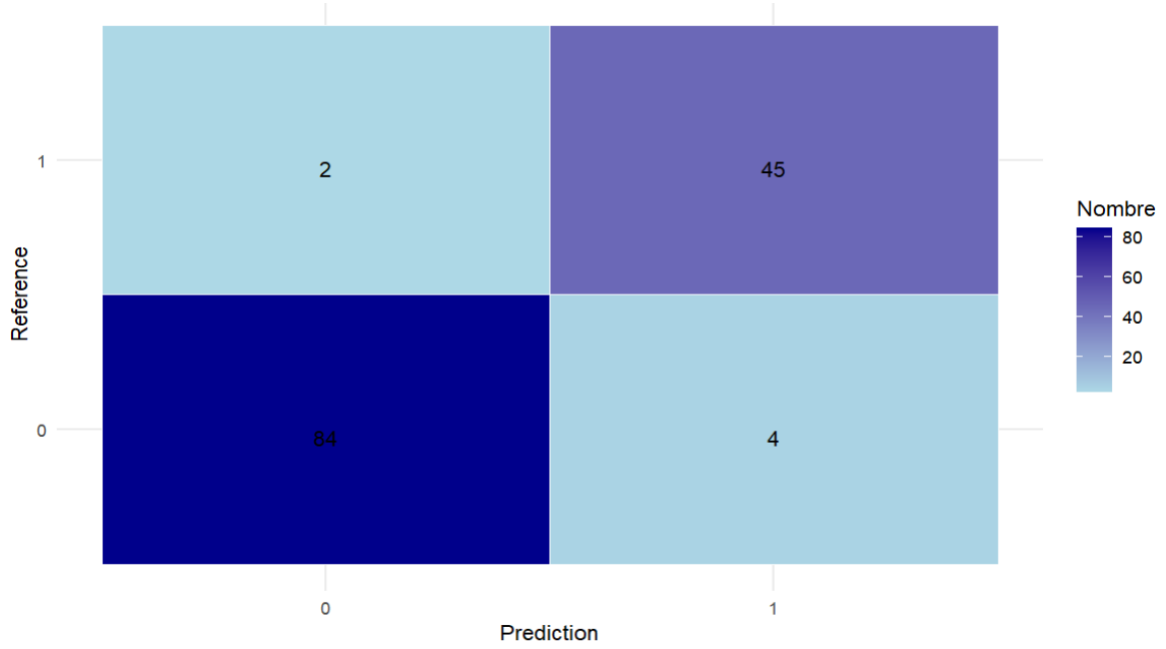
Figure 21 : Importance des variables



Dans le contexte des réseaux de neurones multicouches (MLP), l'importance des variables se révèle cruciale pour comprendre comment ces modèles prennent des décisions. Chaque variable d'entrée peut avoir un impact différent sur les résultats prédits par le MLP. Ainsi, explorer l'importance des variables dans un MLP permet de déterminer quelles caractéristiques ont le plus d'influence sur les prédictions du modèle, offrant ainsi des insights précieux pour interpréter et améliorer sa performance.

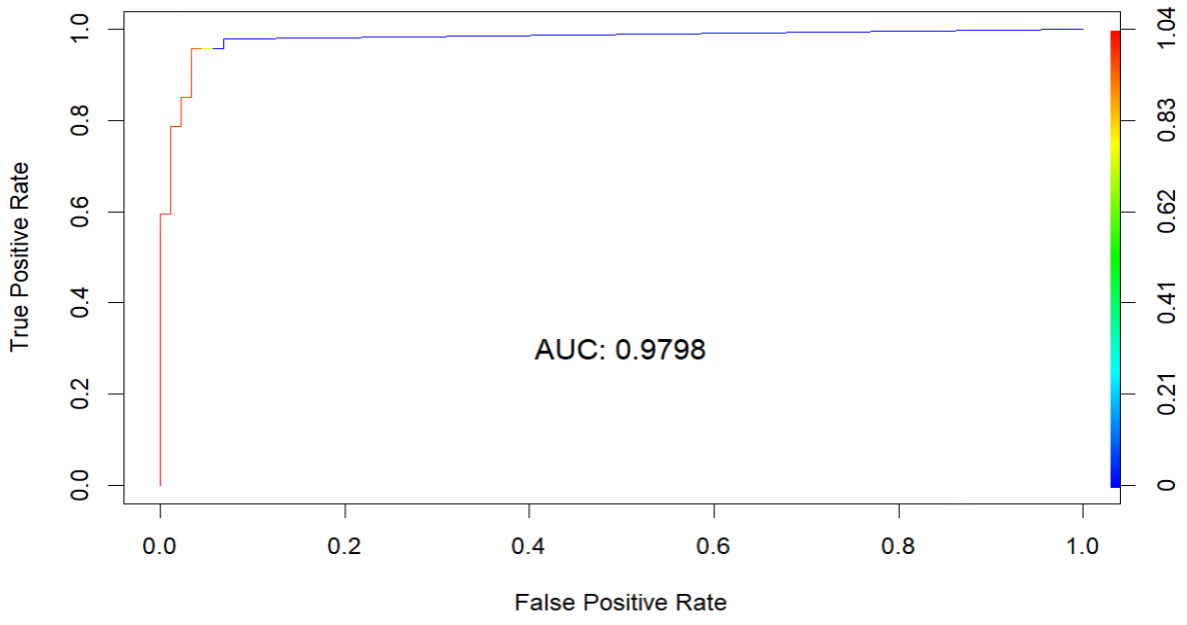
Les résultats de notre analyse indiquent que parmi les trois caractéristiques cellulaires les plus importantes selon le modèle de réseau de neurones, les noyaux nus et l'épaisseur des amas cellulaires sont encore une fois mentionnés en tant que facteurs significatifs.

Figure 22 : Matrice de confusion du jeu test, modèle MLP



La sensibilité, avec un score de 0.9574, met en évidence la capacité du modèle à identifier 95.74% des tumeurs malignes parmi toutes celles qui sont réellement malignes. En parallèle, la spécificité, notée à 0.9545, indique que le modèle a correctement identifié 95.45% des tumeurs bénignes parmi toutes celles qui sont réellement bénignes. Globalement, la précision du modèle s'élève à 0.9556, démontrant qu'il a prédit correctement la classe de 95.56% de toutes les observations.

Figure 23 : Courbe ROC, modèle MLP



5. Conclusion

En conclusion, cette étude visait à examiner l'utilisation et l'adaptabilité de l'intelligence artificielle à des données de santé à l'aide de modèle de machine learning. Dans cette perspective, nous avons analysé un ensemble de données sur les caractéristiques cellulaires de tumeurs du cancer du sein issu d'une étude du Dr William H. Wolberg datant de 1992.

Parmi un large éventail de modèles potentiels, il a été nécessaire de comprendre leur mécanisme et leurs caractéristiques afin de choisir ceux qui s'adaptaient le mieux aux données utilisées. Notre sélection s'est portée sur trois modèles : un modèle de régression logistique, un modèle basé sur les arbres de décisions, et un modèle de réseau de neurones.

Malgré les difficultés rencontrées pour la réalisation et le réglage des hyperparamètres de ces derniers, nous avons développé trois modèles fonctionnels qui offrent une précision très satisfaisante. En effet, nous retrouvons dans l'ordre de performance : la régression logistique (96.3%) puis le réseau de neurones (95.56%) et enfin l'arbre de décisions (93.33%). Il est intéressant de noter que ce n'est pas nécessairement le modèle le plus complexe qui offre les meilleurs résultats, comme en témoigne la performance supérieure de la régression logistique par rapport au réseau de neurones plus sophistiqué.

Au-delà des résultats purement statistiques, cette étude a pu illustrer les pistes d'évolution permises par l'intelligence artificielle pour des enjeux cruciaux tels que le secteur de la santé.

6. Discussion

L'une des principales limites de cette étude est l'utilisation de données datant de 1992. Les méthodes de collecte de données et les technologies médicales ont considérablement évolué depuis cette époque, ce qui pourrait rendre les conclusions de cette étude moins applicables aux données de santé contemporaines. De plus, les données peuvent manquer de certaines caractéristiques importantes qui ne pouvaient pas être mesurées ou enregistrées à l'époque.

Cela nous amène au problème rencontré lors de la recherche des données. En effet, obtenir des données de santé de qualité représente un défi majeur dans le domaine de la recherche. Les données de santé sont souvent sensibles, nécessitant des protocoles stricts pour assurer la confidentialité et la sécurité des informations personnelles. De plus, elles peuvent être fragmentées, issues de différentes sources et formats, rendant leur intégration complexe. Les aspects éthiques et légaux, tels que le respect des réglementations de protection des données (comme le RGPD en Europe), ajoutent une couche de difficulté supplémentaire. Par conséquent, accéder à des ensembles de données suffisamment riches et représentatifs pour entraîner des modèles robustes est une tâche ardue et coûteuse.

Ensuite, les bons résultats obtenus avec les modèles dans cette étude pourraient être en partie attribués à la qualité et à la pertinence des données utilisées. Lorsque les données sont bien structurées et représentatives du phénomène étudié, même des modèles relativement simples peuvent offrir des performances élevées. Cela souligne l'importance de la préparation et de la qualité des données.

Bien que plusieurs modèles de machine learning aient été utilisés, les choix méthodologiques peuvent influencer les résultats. Par exemple, les hyperparamètres des modèles peuvent ne pas avoir été optimisés de manière exhaustive, et d'autres techniques de validation pourraient fournir des résultats différents. De plus, l'étude n'a pas exploré des méthodes de machine learning plus avancées ou récentes, telles que les

réseaux neuronaux profonds ou des forêts aléatoires, qui pourraient potentiellement offrir de meilleures performances.

En conclusion, bien que cette étude ait démontré le potentiel des techniques de machine learning dans l'analyse des données de santé, les limitations mentionnées doivent être prises en compte lors de l'interprétation des résultats et de l'application des conclusions à des contextes réels.

7. Annexe

Figure 1 : Test de Rosner de la variable Mitoses

```
Number of Outliers Detected:     10

   i   Mean.i     SD.i Value Obs.Num    R.i+1 lambda.i+1 Outlier
1  0 1.607407 1.741006    10      64 4.820542   3.941249    TRUE
2  1 1.594955 1.711956    10      70 4.909616   3.940869    TRUE
3  2 1.582467 1.682222    10      84 5.003819   3.940489    TRUE
4  3 1.569940 1.651767    10      97 5.103661   3.940108    TRUE
5  4 1.557377 1.620549    10     162 5.209729   3.939726    TRUE
6  5 1.544776 1.588522    10     182 5.322698   3.939344    TRUE
7  6 1.532138 1.555634    10     229 5.443350   3.938961    TRUE
8  7 1.519461 1.521829    10     231 5.572598   3.938577    TRUE
9  8 1.506747 1.487041    10     273 5.711513   3.938193    TRUE
10 9 1.493994 1.451199    10     286 5.861364   3.937808    TRUE
```

Figure 2 : Normalisation des données

```
preproc_obj <- preProcess(ds[, -which(names(ds) == "class")], method = "range")
```

Le code présenté ci-dessus effectue une normalisation (mise à l'échelle min-max) des données. La méthode utilisée, "range", dans la fonction preProcess de la bibliothèque caret, redimensionne les données de sorte que les valeurs soient comprises entre 0 et 1.

Figure 3 : Suréchantillonnage de l'ensemble d'entrainement

```
training_set_over <- ovun.sample(class ~., data = training_set, method = "over", N = 702,
seed = 123)$data
```

Cette ligne de code utilise la fonction ovun.sample() de la bibliothèque ROSE pour suréchantillonner l'ensemble de données d'entraînement training_set en ajoutant des observations à la classe minoritaire. Plus précisément, cette librairie crée de nouvelles observations pour la classe minoritaire en utilisant des techniques statistiques comme

la duplication aléatoire d'observations existantes ou la génération d'observations synthétiques basées sur les caractéristiques des observations existantes.

Figure 4 : Modèle de régression logistique

```
Call:
glm(formula = class ~ ., family = binomial(link = "logit"), data = training_set_over)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -6.8372     0.8281  -8.256  < 2e-16 ***
clump_thickness     3.8478     1.1940   3.223  0.00127 **
size_uniformity     2.2982     2.6725   0.860  0.38982
shape_uniformity    5.5682     2.6073   2.136  0.03271 *
marginal_adhesion   2.1052     1.1728   1.795  0.07264 .
epithelial_size     1.4540     1.4998   0.969  0.33231
bare_nucleoli       4.4766     0.9969   4.490 7.11e-06 ***
bland_chromatin     1.3054     1.8406   0.709  0.47820
normal_nucleoli     2.3882     1.1654   2.049  0.04044 *
mitoses             5.5926     2.8799   1.942  0.05215 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 973.179  on 701  degrees of freedom
Residual deviance:  95.909  on 692  degrees of freedom
AIC: 115.91
```

8. Bibliographie

[i] https://whatsnext.nuance.com/fr-fr/ai-sante/penurie-de-soignants-et-engorgement-des-urgences-en-france/

[ii] Ng, A.Y., Oberije, C.J.G., Ambrózay, É. et al. Prospective implementation of AI-assisted screen reading to improve early detection of breast cancer. Nat Med 29, 3044–3049 (2023).

[iii] Jacques Rouëssé, Histoire du cancer du sein en occident, enseignements et réflexions, éditions Springer, 2011.

[iv] https://www.has-sante.fr/jcms/c_1741602/fr/cancer-du-sein-quel-depistage-selon-vos-facteurs-de-risque-questions-/-reponses#toc_1_1

[v] https://ishh.fr/cancer-du-sein/facteurs-de-risques-hereditaires-du-cancer-du-sein/

[vi] https://www.science-et-vie.com/corps-et-sante/cancer-du-sein-decouverte-dun-mecanisme-cellulaire-favorisant-la-propagation-des-cellules-tumorales-66748.html

[vii] https://ishh.fr/cancer-du-sein/les-differentes-tumeurs-benignes-du-sein/

[viii] Fan, Y., & Bradley, A. P. (2016). A method for quantitative analysis of clump thickness in cervical cytology slides. Micron, 80, 73-82.

[ix] Choi, S. W., Zhang, Y., & Xia, Y. (2010). Three-dimensional scaffolds for tissue engineering: the importance of uniformity in pore size and structure. Langmuir, 26(24), 19001-19006.

[x] Clark, E. A., & Lane, P. J. (1991). Regulation of human B-cell activation and adhesion. Annual review of immunology, 9(1), 97-127.

[xi] Ma, Y. C., Wang, L., & Yu, F. L. (2013). Recent advances and prospects in the isolation by size of epithelial tumor cells (ISET) methodology. Technology in Cancer Research & Treatment, 12(4), 295-309.

[xii] Bénédicte Royer, Claude Bigorgne, Marie-Annick de Maublanc, Cytopathologie des tumeurs malignes épithéliales, Revue Française des Laboratoires, Volume 2002, Issue 3402002, Pages 39-51, ISSN 0338-9898

[xiii] Sehgal, P., & Chaturvedi, P. (2023). Chromatin and Cancer: Implications of Disrupted Chromatin Organization in Tumorigenesis and Its Diversification. Cancers, 15(2), 466.

[xiv] Shaw, P., & Brown, J. (2012). Nucleoli: composition, function, and dynamics. Plant physiology, 158(1), 44-51.

[xv] Cree, I.A., Tan, P.H., Travis, W.D. et al. Counting mitoses: SI(ze) matters!. Mod Pathol 34, 1651–1657 (2021).

[xvi] Michel Bierlaire, Introduction à l'optimisation différentiable, 2006

[xvii] Rakotomalala, R. (2011). Pratique de la régression logistique. Régression logistique binaire et polytomique, Université Lumière Lyon, 2, 258.

[xviii] Hebiri, M. (2009). Quelques questions de sélection de variables autour de l'estimateur LASSO (Doctoral dissertation, Université Paris-Diderot-Paris VII).

[xix] Caron, S. (2011). Une introduction aux arbres de décision. Stéphane Caron, 31.

[xx] Taud, H., & Mas, J. F. (2018). Multilayer perceptron (MLP). Geomatic approaches for modeling land change scenarios, 451-455.

[xxi] Breast Cancer Wisconsin Dataset. Available at: UCI Machine Learning Repository.

[xxii] Street, W.N., Wolberg, W.H., & Mangasarian, O.L. (1993). Nuclear feature extraction for breast tumor diagnosis. Electronic imaging.

[xxiii] Zhou, W., & Xie, Y. (2013). Interactive medical image segmentation using snake and multiscale curve editing. Computational and mathematical methods in medicine, 2013, 325903.

Table des matières