UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ À
L'UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

COMME EXIGENCE PARTIELLE À L'OBTENTION DE LA MAÎTRISE
AVEC MÉMOIRE EN MATHÉMATIQUES ET INFORMATIQUE
APPLIQUÉES.

PAR
KATIENEFOA SORO

Modèles prédictifs par apprentissage automatique pour la survie : application à la base de données cliniques et génétique pour les tumeurs du cancer du sein.

Juin 2025

RÉSUMÉ

Le cancer du sein est une maladie courante qui touche principalement les femmes, causée par la perturbation de certaines cellules qui se multiplient souvent et forment une masse appelée tumeur. Dans la plupart des cas, le développement du cancer du sein prend plusieurs mois voire plusieurs années.

Récemment, des techniques d'apprentissage automatique (ML) ont été utilisées en biomédecine et en informatique pour prendre des décisions en termes de diagnostic et d'analyse afin de lutter contre le cancer du sein. Ce travail de recherche propose une étude comparative des approches de sélection de caractéristiques et d'apprentissage automatique pour prédire la survie des patientes atteintes d'un cancer du sein.

Ce type de cancer (invasif) : la tumeur cancéreuse peut briser la membrane du tissu d'origine. Des cellules cancéreuses ont alors quitté leur tissu d'origine et se sont localisées dans des tissus voisins.

Pour ce faire, nous allons comparer les performances entre différents algorithmes d'apprentissage automatique : La régression linéaire, Support Vector Machine (SVM), Forêt Aléatoire (RF), Lasso et Ridge dans des ensembles de jeux de données sur le cancer du sein accessibles au public.

Pour ce projet de recherche nous avions deux types de données: Données (Discovery Data) utilisées pour la recherche fondamentale afin d'identifier de nouveaux biomarqueurs, des mutations génétiques ou des voies moléculaires impliquées dans le cancer du sein et ensuite nous avons les données (Clinical Data) centrées sur le patient recueilli pendant le diagnostic, le traitement et les suivis pour évaluer la progression de la maladie, la réponse au traitement et les résultats de survie.

L'objectif principal est d'évaluer l'exactitude de l'apprentissage automatique sur les données par rapport à l'efficacité de chaque algorithme en termes, de précision, de sensibilité et de spécificité. Les résultats expérimentaux montrent que la Forêt aléatoire donne la plus grande précision (90%).

ABSTRACT

Breast cancer is a common disease that affects women, caused by the disruption of certain cells that often multiply and form a mass called a tumor. In most cases, the development of breast cancer takes several months or even years. Recently, machine learning (ML) techniques have been used in biomedicine and informatics to make diagnostic and analytical decisions to fight breast cancer. This research proposes a comparative study of trait selection and machine learning approaches to predict survival in patients with grade 3 breast cancer. This type of cancer is stage 3 (invasive): the cancerous tumor can break the membrane of the original tissue. Cancer cells then left their original tissue and localized themselves to nearby tissues. To do this, we compared the performance between different machine learning algorithms: Linear Regression, Support Vector Machine (SVM), Random Forest (RF), Lasso and Ridge in publicly available Discovery Data sets.

For this research project we had two types of data: Discovery Data used for basic research to identify new biomarkers, genetic mutations or molecular pathways involved in breast cancer and Patient-centered Clinical Data collected during diagnosis, treatment and follow-ups to assess disease progression, response to treatment and survival outcomes. The main goal is to evaluate the accuracy of data machine learning against the efficiency of each algorithm in terms of accuracy, precision, sensitivity, and specificity. The experimental results show that Random Forest gives the highest accuracy (90%).

AVANT-PROPOS

Le cancer du sein demeure l'une des principales causes de mortalité chez la femme dans le monde. La complexité de cette maladie, liée à une grande hétérogénéité biologique, nécessite des outils prédictifs performants pour optimiser la prise en charge clinique.
L'objectif principal de ce travail est de développer des modèles prédictifs robustes et fiables destinés aux professionnels de santé, leur permettant d'orienter les décisions thérapeutiques et le suivi des patientes atteintes de cancer du sein.

Pour cela, nous exploitons un ensemble de données intégrant à la fois des informations cliniques et génétiques. Afin d'améliorer la pertinence et la performance des modèles, des méthodes de sélection de caractéristiques sont appliquées, notamment la corrélation de Pearson, la corrélation de Spearman et l'analyse par information mutuelle. Ces techniques permettent d'identifier les variables les plus informatives en lien avec la survie des patientes.

Ensuite, plusieurs algorithmes d'apprentissage automatique sont mis en œuvre pour construire les modèles prédictifs : la forêt aléatoire, la régression linéaire, les régressions régulières (Ridge, Lasso, Elastic Net) ainsi que les modèles de régression à vecteurs de support (SVR) dans leurs variantes linéaires et non linéaires.

Ce mémoire présente ainsi une approche intégrée combinant sélection de caractéristiques et apprentissage automatique, visant à fournir des outils d'aide à la décision cliniques plus précis pour la gestion du cancer du sein.

DÉDICACES

À toutes les femmes courageuses qui luttent contre le cancer du sein, à celles qui ont combattu et qui continuent de combattre avec une force et une détermination incomparable. Que ce mémoire soit un hommage à votre résilience, à votre dignité et à votre espoir.

Vous êtes une source d'inspiration et de courage pour nous tous.

Je dédie ce travail à toutes les femmes qui, malgré les difficultés et les épreuves, n'ont jamais perdu espoir. À celles qui nous ont quittés, dont le souvenir restera à jamais dans nos cœurs, et à celles qui se battent aujourd'hui, avec l'espoir de voir un jour cette maladie éradiquée.

Que la recherche, la solidarité et l'amour vous accompagnent dans chaque étape de votre parcours.

REMERCIEMENTS

L’écrivaine Madeleine Ferron déclarait dans son œuvre le chemin des dames : « Dans la vie, les hommes sont tributaires les uns des autres. Il y a donc toujours quelqu'un à maudire ou à remercier »**[1].**

C’est donc avec gratitude que je remercie toutes les personnes qui ont contribué de près ou de loin à la réalisation de ce mémoire. Mais au-delà de cela, je tiens à exprimer ma reconnaissance envers celles et ceux qui ont rendu ce projet possible ici, à l’Université du Québec à Trois-Rivières.

Je remercie tout d'abord les membres du jury, pour l'honneur qu'ils m'ont fait en évaluant mon travail, ainsi que l'ensemble du corps professoral et administratif de l'Université du Québec à Trois-Rivières pour leur soutien. Je tiens particulièrement à remercier Madame Nadia Ghazzali et Monsieur Venkata Manem pour leur encadrement pédagogique constant et leur suivi rigoureux tout au long de l'avancement de mon projet. Leur expertise et leur disponibilité ont été précieuses dans la conduite de cette étude. Enfin, je ne peux clore cette série de remerciements sans penser à mes parents qui ont été d’une grande aide et Madame Jane Alice Van Den Berg , qui a fait de mes études une priorité et m'a permis d'en arriver là aujourd’hui.

Je remercie très particulièrement les enseignants du département de Mathématiques et informatique appliquées, aussi ma promotion 2022-2023 de Master en mathématiques et informatique appliquées et toutes les personnes qui ont soutenues de près ou de loin dans la Réalisation de ce travail.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

TABLE DES FIGURES

INTRODUCTION GÉNÉRALE

Introduction

Le cancer du sein est l'un des problèmes de santé publique les plus fréquents au monde, en particulier chez les femmes.

Il affecte principalement les femmes âgées de 50 à 70 ans, bien qu'il puisse aussi, dans de rares cas, affecte soit 1% des hommes.

Cette maladie se caractérise par une croissance incontrôlée de cellules anormales dans le tissu mammaire.

Cette prolifération cellulaire entraîne la formation d'une tumeur cancéreuse, un amas de cellules pouvant envahir et détruire les tissus environnants.

Le cancer du sein est une maladie multifactorielle, pouvant évoluer sur plusieurs mois ou années.

En général, son développement est dû à des mutations génétiques qui rendent les cellules cancéreuses défectueuses.

Il s'agit d'une maladie très hétérogène, avec des formes et des évolutions variables selon les individus.

Statistiques Clés sur le Cancer du Sein :

📌 Au Canada

✓ Le cancer du sein est le cancer le plus fréquent chez les femmes.

✓ Il représente la 2e cause de décès par cancer chez les Canadiennes.

✓ Environ 1 % des cas concernent des hommes.

📌 Au Québec en 2022

✓ 8 324 femmes ont reçu un diagnostic de cancer du sein.

✓ Les femmes de 50 à 69 ans sont les plus touchées.

✓ 78 hommes ont également été diagnostiqués.

📉 Tendances et Risques.

- 1 femme sur 8 développera un cancer du sein au cours de sa vie.
- 1 femme sur 36 en mourra.

Programme québécois de dépistage du cancer du sein (*PQDCS*) **[2].**

Ce programme a pour objectif de détecter la maladie à un stade précoce afin d'améliorer les chances de traitement et de survie.

Ces dernières années, le domaine de L'intelligence artificielle et les techniques d'apprentissage automatique ont joué un rôle majeur dans des domaines médicaux tels que la détection assistée par ordinateur. Ce travail combine une approche traditionnelle d'apprentissage machine avec une approche d'apprentissage profonde pour augmenter l'efficacité de la détection du cancer de sein tout comme le *PQDCS*. Les techniques d'apprentissage en profondeur peuvent s'avérer un outil informatique utile pour modéliser le comportement de l'expert, améliorer la précision de la prédiction de la survie et devenir une norme universelle pour les médecins.

Motivation et énoncé du problème

Le diagnostic du cancer du sein repose sur une série d'examens cliniques, biologiques et d'imagerie, souvent coûteux et nécessitant des infrastructures spécialisées. Dans l'esprit de beaucoup, aucun diagnostic de cancer n'est plus redouté que celui-ci.

Le cancer est souvent considéré comme une maladie incurable non diagnostiquée et incroyablement douloureuse. Cette vision effrayante peut être exagérée, il ne fait aucun doute que le cancer est une maladie grave, mais la réalité est que divers cancers peuvent aujourd'hui être traités, éliminés ou ralentis. Avec les progrès scientifiques en matière de détection et de diagnostic, il y a plus d'espoir que de désespoir dans la plupart des cas de cancer.

Ces dernières années, les avancées en intelligence artificielle ont permis d'améliorer la précision et la rapidité des diagnostics, tout en facilitant la prise de décision clinique et en contribuant à de meilleurs résultats en matière de santé. Ce projet vise à combiner une approche de sélection de caractéristiques avec des techniques d'apprentissage automatique pour prédire la survie des patientes atteintes d'un cancer du sein de stade 3, en s'appuyant

sur une base de données utilisée dans la recherche fondamentale. L'objectif est d'identifier de nouveaux biomarqueurs, des mutations génétiques ou des voies moléculaires impliquées dans le développement du cancer du sein.

Questions de recherche

Notre travail consiste à analyser et comparer plusieurs méthodes d'apprentissage automatique, identifiées au cours de notre étude de l'état de l'art dans le domaine de la détection du cancer du sein. Les algorithmes étudiés incluent la régression linéaire (LR), les machines à vecteurs de support (SVM), les forêts aléatoires (Random Forest), ainsi que les régressions pénalisées Lasso et Ridge.

Nous utilisons différentes caractéristiques issues des données pour évaluer et comparer les performances de ces techniques. L'objectif est de déterminer les modèles les plus efficaces dans le cadre de la classification appliquée à la détection du cancer du sein. Une analyse comparative est menée sur la base de plusieurs métriques de performance, suivie d'une discussion approfondie des résultats obtenus.

Quel est l'algorithme d'apprentissage supervisé le plus performant et le plus fiable pour prédire la survie des patients atteints du cancer du sein, en utilisant des caractéristiques extraites des données cliniques ou biologiques ?

Cette étude cherche à répondre aux interrogations qui circulent au sein de l'organisation.

Structure du mémoire

Le mémoire est structuré comme suit :

Le chapitre 1 :C'est une présentation des généralités sur le cancer de sein.

Le chapitre 2 : On présente les méthodes de sélection de caractéristiques et d'apprentissage automatique pour la prédiction de la survie du cancer du sein.

Le chapitre 3 : Description de la base de données utilisée et les techniques de prétraitement que nous avons utilisées suivi par le résultat final.

Une conclusion et des perspectives concluent ce mémoire.

CHAPITRE 1 : GÉNÉRALITÉ SUR LE CANCER

Introduction générale

Le Cancer contient un ensemble de maladies qui se caractérisent par la multiplication et la propagation anarchique de cellules anormales. Si les cellules cancéreuses ne sont pas éliminées, l'évolution de la maladie va mener plus ou moins rapidement au décès de la personne touchée.

Une tumeur est une masse formée de cellules qui peuvent être malignes ou bénignes. Parmi ces cancers, le cancer du sein est devenu un problème de santé publique majeur avec une réelle urgence d'intervention et de prise en charge.

Au Canada en l'occurrence, on assiste à une véritable transition épidémiologique marquée par l'amorce de la transition démographique, l'augmentation de l'espérance de vie, la transformation de l'environnement et les changements de mode de vie. D'ailleurs, depuis quelques années le cancer du sein est devenu un véritable problème de santé publique dans tout le pays.

Le 13 octobre est spécifiquement reconnu comme la Journée nationale du cancer du sein métastatique, visant à mettre en lumière cette forme avancée de la maladie et à souligner l'importance de la recherche dans ce domaine. De plus, le 19 octobre est célébré comme la Journée mondiale de lutte contre le cancer du sein, avec des événements organisés pour honorer les personnes touchées et promouvoir la recherche.

1.1 Anatomie du sein

Le sein est composé d'une **glande mammaire**, de fibres de soutien (ligaments de Cooper) et de graisse (tissu adipeux); le tout est recouvert par la peau. La quantité de chacune de ses composantes peut varier d'une femme à l'autre.

Chez l'homme, le sein se compose aussi de tissu graisseux, de canaux et de lobules, mais qui sont en moins grand nombre que chez la femme. Le sein est situé par-dessus le muscle pectoral. On trouve également dans le sein, des nerfs, des vaisseaux sanguins et lymphatiques.

La **glande mammaire** est divisée en 15 à 20 sections qu'on appelle **lobes,** composés de **lobules**. Ceux-ci sont reliés à des **canaux** qui se rendent sous le mamelon (situé au centre du sein). On peut également observer des chaînes de ganglions lymphatiques qui filtrent les microbes et protègent le corps contre l'infection et la maladie. Le cancer du sein peut se développer tant au niveau d'un canal galactophore que d'un lobule et il peut également se retrouver au niveau des ganglions lymphatiques.

Chez l'homme, le cancer est principalement présent dans les canaux et beaucoup plus rarement dans les lobules.

1.2 Qu'est-ce-que le cancer du sein ?

Le cancer du sein est une maladie caractérisée par la croissance incontrôlée de cellules anormales dans les tissus mammaires. Ces cellules peuvent former une tumeur qui, si elle n'est pas traitée, peut envahir les tissus environnants et se propager à d'autres parties du corps, provoquant des métastases.

Bien que les hommes puissent également développer un cancer du sein, cette maladie touche principalement les femmes. En effet, les seins féminins contiennent des glandes mammaires, responsables de la production de lait maternel pour l'allaitement des nouveau-nés. C'est pourquoi les seins sont considérés comme des organes accessoires de l'appareil reproducteur féminin.

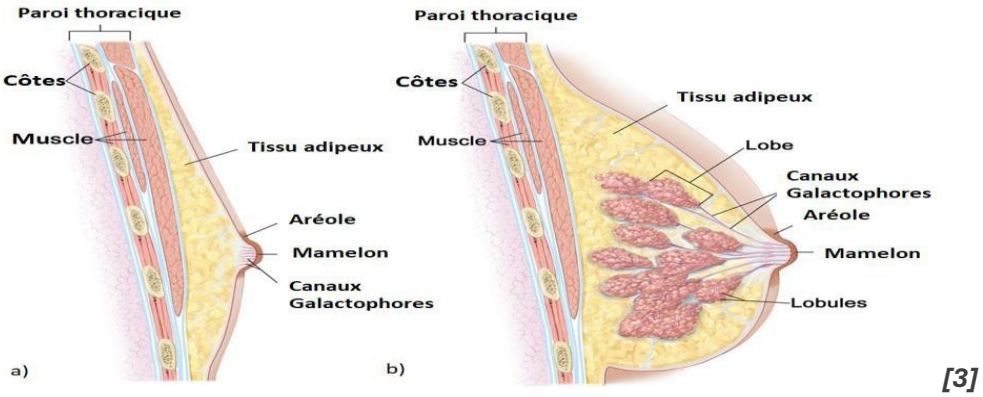


Figure 1-1 : Anatomie du sein a) chez l'homme et b) chez la femme

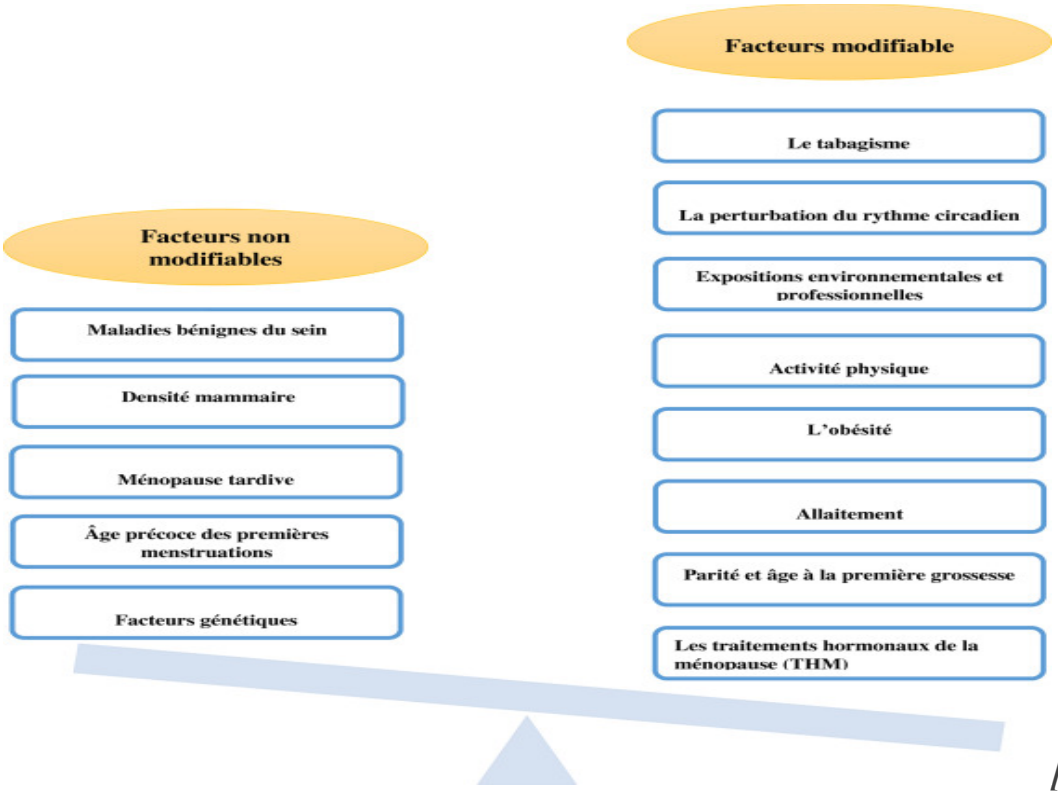
Les seins occupent la partie antérosupérieure du thorax en avant des muscles pectoraux. Ils ne contiennent pas de muscles et sont soutenus par des ligaments.
Le sein correspond à l'organe qui appartient à la **glande mammaire (glande exocrine)** qui se développe à partir de la puberté chez la femme.
Le sein est fait de (graisse, tissu conjonctif, glandes et de canaux).
***1-La glande mammaire* :** Est une masse de densité variable, organisée en une vingtaine de lobes. Les glandes produisent du lait quand elles sont stimulées par les hormones de la femme en cours de grossesse.

2- Les lobes* :** Sont des groupes de glandes qui produisent le lait, Ils sont séparés et maintenus par du tissu « ***conjonctif et adipeux ». Chaque sein comporte de 15 à 25 lobules. (Figure1.2)
***3-Les ligaments* :** Sont des bandes serrées de tissu conjonctif qui soutiennent les seins. Ils traversent le sein de la peau jusqu'aux muscles où ils se fixent au thorax.
***4-Les canaux* :** Sont des tubes qui transportent le lait des lobules au mamelon.
***5-Le mamelon* :** Est la région située au centre de l'aréole et d'où sort le lait à une extrémité. Le mamelon est fait de fibres musculaires. Quand ces fibres se contractent, le mamelon durcit, ou pointe vers l'extérieur.
***6-L'aréole* :** Est la surface ronde, rosée ou brunâtre qui entoure le mamelon. Elle contient de petites glandes qui libèrent, ou sécrètent, une substance huileuse qui agit comme lubrifiant pour le mamelon et l'aréole.

1.3 Les causes et les facteurs de risques

Les causes exactes du cancer du sein sont complexes, mais plusieurs facteurs de risque augmentent la probabilité de développer la maladie. Les plus importants ne sont pas modifiables :
***Facteur Age* :** Le risque augmente sensiblement après 50 ans, mais la femme plus jeune peut être également touchée.
***Antécédents familiaux* :** les femmes dont les sœurs, les mères ou les filles ont un cancer du sein présentent un risque plus important. En particulier si les membres de la famille sont atteints avant l'âge de 50 ans.
Facteurs génétiques (Mutations des gènes BRCA1 et BRCA2): environ 5 à 10 % de tous les cancers du sein sont déclenchés par une prédisposition héréditaire. Les femmes concernées sont souvent malades avant 50 ans.
***Facteurs hormonaux* :** le risque de cancer du sein est légèrement plus élevé chez les femmes dont la première menstruation s'est produite avant l'âge de 12 ans, dont la dernière a eu lieu après l'âge de 55 ans, chez les femmes qui n'ont pas d'enfants ou qui ont accouché après l'âge de 30 ans. Le mode de vie joue également un rôle. Les facteurs suivants peuvent légèrement augmenter le risque:

Le traitement hormonal pour les troubles liés à la « ménopause, la prise de la pilule contraceptive, le tabagisme, la consommation excessive d'alcool, l'obésité, une alimentation déséquilibrée, riche en lipides, le manque d'activité physique ».



[4].

Figure 1-2 : les facteurs de risque du cancer du sein.

1.4 Les symptômes

Le médecin spécialiste doit absolument vous examiner si vous avez de tels symptômes. La plus-part du temps, les femmes touchées découvrent elles - mêmes au niveau de la poitrine quelque chose d'anormal. Dans neuf cas sur dix, ces troubles ne sont pas dus à un cancer.

Le cancer du sein peut être asymptomatique à ses débuts. Certains signes doivent alerter :

- ✅ Rougeur, peau d'orange ou ulcération du sein.
- ✅ Masse ou nodule dans le sein ou l'aisselle (douloureux ou non).
- ✅ Modification de la forme ou de la taille du sein.
- ✅ Rétraction ou écoulement anormal du mamelon.

1.5 Diagnostic et dépistage

Pour diagnostiquer un cancer du sein, on a recours en premier lieu à des procédés.

La détection précoce améliore les chances de guérison. Les principaux examens sont :

- Mammographie : Examen radiologique clé pour le dépistage.
- Échographie mammaire : Complète la mammographie en cas de doute.
- Biopsie : Confirme le diagnostic en analysant un prélèvement de tissu.
- IRM mammaire : Utile pour les femmes à haut risque.

La mammographie (radiographie du sein) : est une radiographie à faible dose du sein. L'image obtenue est appelée cliché mammaire. Elle peut aider à détecter des tumeurs cancéreuses (malignes) et des tumeurs non cancéreuses (bénignes) dans le sein. Les images communiquent des informations sur la nature, l'emplacement et la taille d'un nodule.

La biopsie (prélèvement d'un échantillon de tissu) : Lors de la biopsie, le médecin prélève à l'aide d'une aiguille ou d'un trocart à biopsie (instrument médical en forme de poinçon) des échantillons de tissu du nodule suspect, que l'on analyse ensuite au microscope.

Parfois par une IRM (L'imagerie par résonance magnétique) **[5].**

1.6 Les traitements

Si vous avez découvert un cancer du sein, l'équipe soignante décident d'entamer un traitement, il existe plusieurs moyens pour traiter le cancer du sein.

Le traitement dépend du type, du stade et des caractéristiques biologiques de la tumeur.

Chirurgie : Les types de chirurgie qu'on vous proposera dépendront surtout des facteurs suivants :

- Taille et emplacement de la tumeur.
- Taille du sein atteint.
- Propagation du cancer aux ganglions lymphatiques.
- Traitements déjà reçus pour le cancer du sein.

Radiothérapie : Lors de la radiothérapie externe, on a recours à un appareil pour diriger la radiation à travers la peau vers la tumeur et le tissu qui l'entoure.

Chimiothérapie : La chimiothérapie est un traitement courant du cancer du sein.

On l'administre souvent après la chirurgie d'un cancer du sein précoce afin de réduire le risque de réapparition de la maladie.

Hormonothérapie : On administre souvent une hormonothérapie pour traiter le cancer du sein dont les récepteurs hormonaux sont positifs. Les femmes ménopausées reçoivent des médicaments hormonaux différents de ceux qu'on administre aux femmes pré-ménopausées.

Immunothérapie : L'immunothérapie aide à renforcer ou à rétablir la capacité du système immunitaire à combattre le cancer. On l'appelle parfois thérapie biologique. On peut administrer une immunothérapie pour :

- Détruire les cellules du cancer du sein.
- Interrompre la croissance et la propagation des cellules cancéreuses du sein.
- Maitriser les symptômes du cancer du sein métastatique.

Thérapies ciblées : Attaquent des molécules spécifiques du cancer (ex : HER2).

1.7 Conclusion

Dans ce chapitre, on a donné les généralités sur le cancer du sein dans lequel s'inscrit le cadre de ce mémoire. Nous avons parlé du cancer en tant que maladie, puis nous avons abordé les facteurs de risque du cancer du sein, les moyens de dépistage de la maladie et les symptômes qu'elle peut provoquer chez la personne touchée. Ensuite, nous nous sommes intéressés au diagnostic et le traitement de cette maladie.

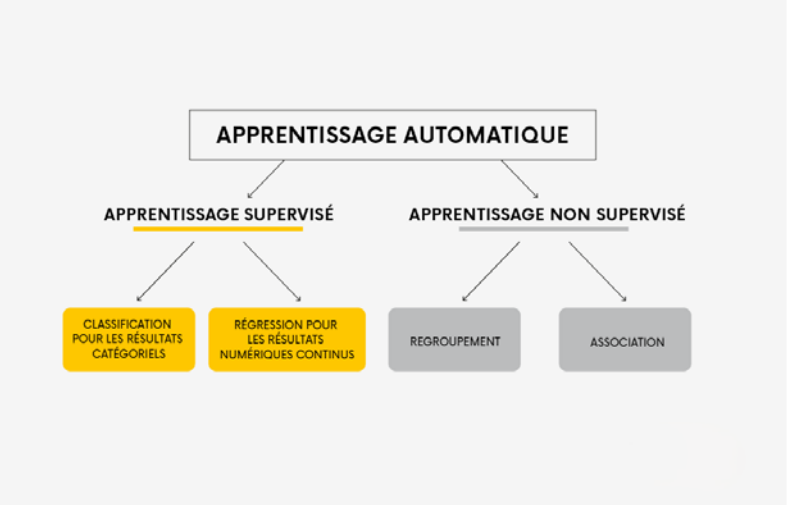
Le cancer du sein est une maladie grave mais détectable et traitable. Un dépistage précoce améliore considérablement les chances de guérison.

CHAPITRE 2 : APPRENTISSAGE AUTOMATIQUE

Introduction

Ce chapitre donne un aperçu des bases théoriques pertinents à l'apprentissage automatique, on fait un état de l'art qui donne une vision générale sur les méthodes de classification, de régression et des approches de sélection de caractéristiques.

L'apprentissage automatique (*ML* ou *Machine Learning*) est une branche de l'intelligence artificielle qui permet aux ordinateurs d'apprendre à partir de données sans être explicitement programmés. Il repose sur des algorithmes qui analysent des ensembles de données pour identifier des motifs et prendre des décisions ou faire des prédictions. Ainsi, le but essentiel de ML est de construire la relation entre les objets et leurs catégories pour la prédiction et la découverte des connaissances **[6]**. Nous allons à présent décrire les principaux algorithmes du Machine Learning et les méthodes de sélection de caractéristiques qui vont nous permettre de réaliser ce projet. Les principaux types d'apprentissage automatique sont : « l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage par renforcement ». Dans ce projet, nous n'aborderons que l'apprentissage supervisé.

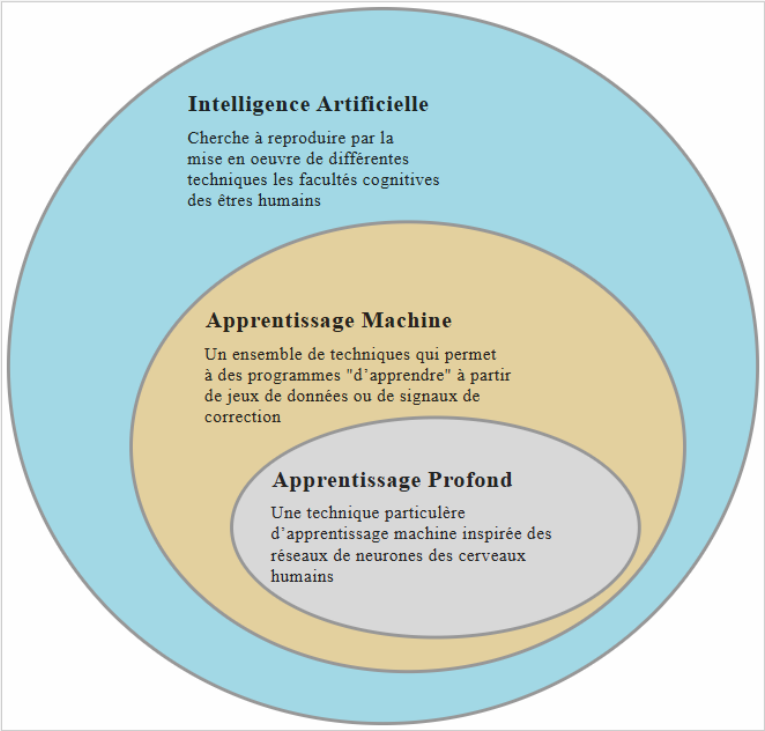


[7].

Figure 2-1 : Taxonomie de l'apprentissage Automatique

2.1 Les concepts et les terminologies de base

L'apprentissage automatique est un des champs d'étude de l'intelligence artificielle. Il fait référence à la capacité d'un système à acquérir et intégrer de façon autonome des connaissances **[8].** L'apprentissage automatique fait référence au développement, l'analyse et l'implémentation de méthodes qui permettent à une machine (au sens large) d'évoluer et de remplir des tâches associées à une intelligence artificielle grâce à un processus d'apprentissage. Cet apprentissage permet d'avoir un système qui s'optimise en fonction de l'environnement, les expériences et les résultats observés **[9].** Les méthodes de l'apprentissage automatique forment une classe de techniques attrayantes pour l'accomplissement des tâches d'extraction de connaissances évoquées. Bien choisis, ces outils peuvent être amenés à accompagner, voire à remplacer l'opérateur humain.



[10].

Figure 2-2 : Représentation des relations entre intelligence artificielle, apprentissage automatique et apprentissage profond.

2.2 Machine learning supervisé

L'approche de l'apprentissage supervisé en ML consiste à utiliser des jeux de données étiquetés qui entraînent des algorithmes pour classer les données ou prédire des résultats avec précision. Le modèle exploite les données étiquetées pour mesurer la pertinence des différentes caractéristiques afin d'affiner progressivement l'ajustement du modèle en fonction du résultat connu.

Il existe deux grandes catégories d'apprentissage supervisé :

- **Classification**
- **Régression**

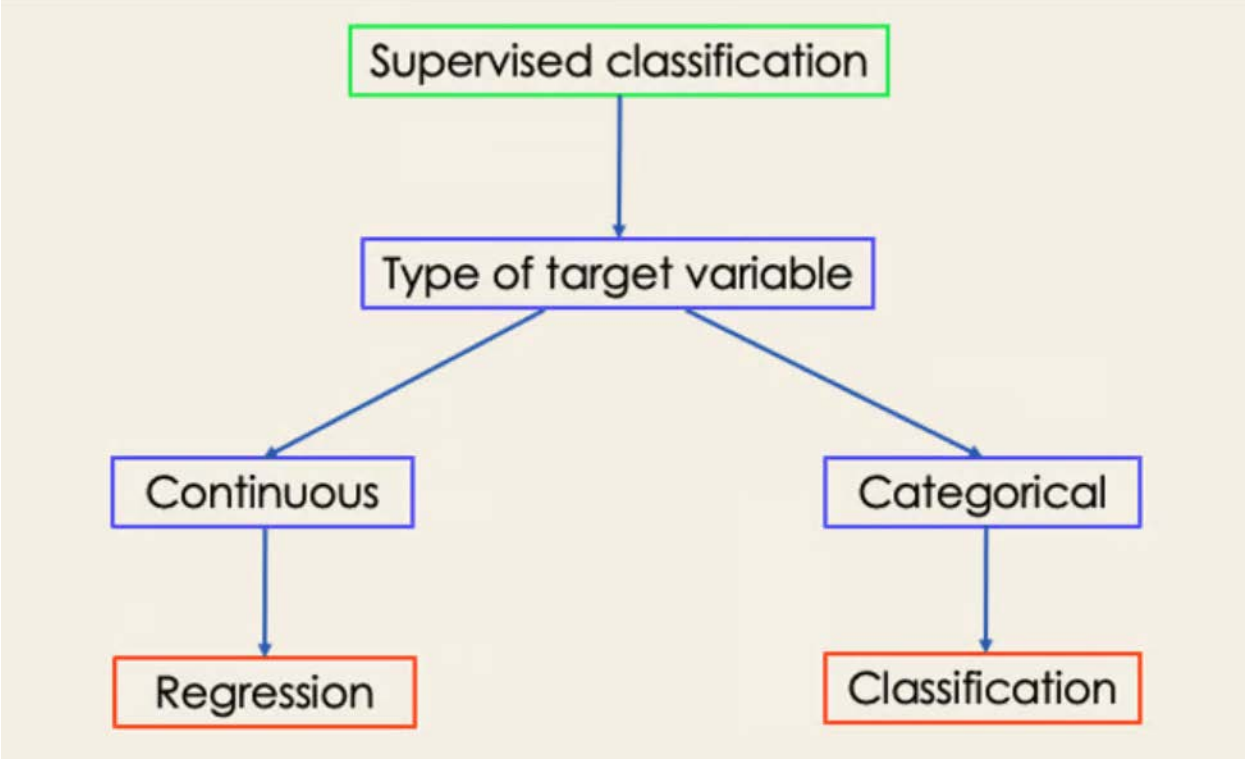


Figure 2-3 : Les processus d'apprentissage automatique supervisé.

2.2.1 La classification

La classification est une méthode d'apprentissage automatique supervisé où le modèle tente de prédire l'étiquette correcte d'une donnée d'entrée donnée. Dans ce cas, le modèle est entièrement entraîné à partir des données d'entraînement, puis évalué sur des données de test avant d'être utilisé pour effectuer des prédictions sur de nouvelles données non vues. L'objectif principal de la classification est d'assigner une observation à une ou plusieurs catégories (ou classes) en fonction de ses caractéristiques ou attributs. Dans un problème de classification, l'ensemble de données est composé de plusieurs exemples, chacun avec un ensemble de caractéristique**s** (ou variables indépendantes) et une étiquette ou classe associée (la variable dépendante).

L'objectif est de construire un modèle capable de prédire l'étiquette ou la classe de nouvelles observations en fonction des caractéristiques qu'elles possèdent. Il existe plusieurs types de classification, en fonction de la nature des classes ou des données : Nous avions par exemple la Classification binaire : Il y a seulement deux classes possibles.

Par exemple, prédire si un courriel est un spam ou non, ou si un patient est atteint d'une maladie ou non. Les classes sont souvent notées comme 0 et 1, il y'a aussi la classification multi class : Il existe plus de deux classes possibles. Par exemple, classer des images en plusieurs catégories : chat, chien, oiseau, etc. et la classification multi label : Une observation peut appartenir à plusieurs classes en même temps. Par exemple, un film peut appartenir simultanément aux classes action, aventure, science-fiction.

Il existe de nombreuses techniques de classification, parmi lesquelles les plus courantes sont : les Arbres de décision : Un arbre qui découpe l'espace des caractéristiques en fonction de tests sur les variables, pour assigner des classes aux observations. L'arbre est construit en choisissant les meilleures variables et seuils qui séparent au mieux les classes, les Machines à vecteurs de support (SVM) : Un algorithme qui cherche à maximiser la séparation entre les classes en trouvant un hyperplan optimal. Il est particulièrement efficace dans des espaces de grande dimension et dans des problèmes non linéaires, les K-plus proches voisins (K-NN) : Un algorithme basé sur la proximité. L'idée est d'assigner une classe à une observation en fonction de la classe des K voisins les plus proches dans l'espace des caractéristiques, la Régression logistique : Bien qu'il porte le nom de régression, il s'agit d'un modèle de

classification utilisé principalement pour les problèmes binaires. La régression logistique calcule la probabilité qu'une observation appartienne à une classe donnée, les Forêts aléatoires (Random Forest) : Un ensemble d'arbres de décision qui combine plusieurs arbres pour améliorer la précision et réduire le risque de sur-apprentissage (overfitting).

La classification est utilisée dans de nombreux domaines, tels que : la Reconnaissance d'image et de vidéo : Par exemple, reconnaître des objets dans des images (voitures, visages, etc.), le Traitement du langage naturel (NLP) : Identifier le sentiment dans un texte (positif, négatif, neutre), ou classer des documents dans différentes catégories (par exemple, classer des articles de news selon les thèmes : politique, sport, etc.), la Médecine : Diagnostiquer des maladies à partir de données médicales, comme classifier des patients en fonction de leurs symptômes (par exemple, détection du cancer etc.).

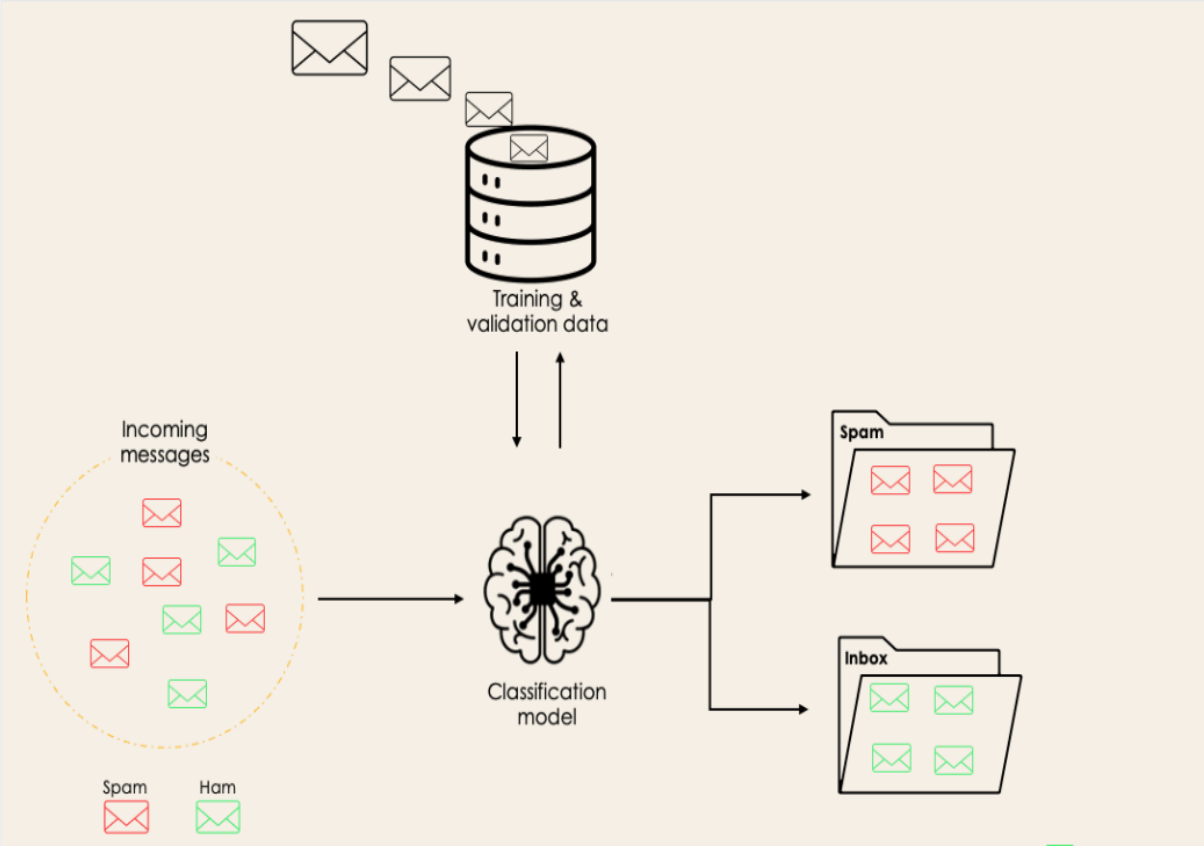


Figure 2-4 : Exemple de classification, si un courriel donné est un spam ou un Ham (pas de spam).

2.2.2 La régression

La régression, comme nous l'avions mentionné (Figure 2-3), est une technique d'apprentissage supervisé utilisée pour prédire des valeurs continues. En d'autres termes, elle permet de modéliser la relation entre une variable dépendante (la variable à prédire) et une ou plusieurs variables indépendantes (les caractéristiques ou facteurs influençant la variable dépendante). La régression est utilisée dans de nombreux domaines similaires à ceux des classificateurs (comme la classification), mais contrairement à ces derniers qui prédisent des classes discrètes, la régression prédit des valeurs continues. Par exemple, la régression a été utilisée dans des études **[11],** où des modèles de régression ont été appliqués pour prédire des scores continus indiquant le niveau de fonctionnement des individus à haut risque de psychose. Ces scores peuvent être utilisés pour évaluer des aspects du fonctionnement quotidien des individus.

En pratique, les modèles de régression sont utilisés pour une variété de tâches dans des domaines comme la prévision des prix immobiliers (par exemple, prédire la valeur d'une maison en fonction de ses caractéristiques), les prévisions de ventes (estimer les ventes futures d'un produit), les résultats d'examen (prédire les résultats d'un étudiant en fonction de ses performances passées et d'autres facteurs) et aussi les mouvements boursiers (prédire l'évolution des prix des actions en fonction de facteurs économiques, financiers, etc.).

2.3 Machine learning non supervisé

Dans le domaine informatique et de l'intelligence artificielle, l'apprentissage non supervisé désigne la situation d'apprentissage automatique où les données ne sont pas étiquetées (par exemple étiquetées comme « balle » ou « poisson »). Il s'agit donc de découvrir les structures sous-jacentes à ces données non étiquetées. Puisque les données ne sont pas étiquetées, il est impossible à l'algorithme de calculer de façon certaine un score de réussite. Ainsi, les méthodes non supervisées présentent une auto-organisation qui capture les modèles comme des densités de probabilité ou, dans le cas des réseaux de neurones, comme combinaison de

préférences de caractéristiques neuronales encodées dans les poids et les activations de la machine. Les autres niveaux du spectre de supervision sont l'apprentissage par renforcement où la machine ne reçoit qu'un score de performance numérique comme guide, et l'apprentissage semi-supervisé où une petite partie des données est étiquetée. L'introduction dans un système d'une approche d'apprentissage non supervisé est un moyen d'expérimenter l'intelligence artificielle. En général, des systèmes d'apprentissage non supervisé permettent d'exécuter des tâches plus complexes que les systèmes d'apprentissage supervisé, mais ils peuvent aussi être plus imprévisibles. Même si un système d'IA d'apprentissage non supervisé parvient tout seul, par exemple, à faire le tri entre des chats et des chiens, il peut aussi ajouter des catégories inattendues et non désirées, et classer des races inhabituelles, introduisant plus de bruit que d'ordre.

2.3.1 Clustering non supervisé

Le clustering non supervisé est une tâche qui, sans supervision, vise à regrouper un ensemble d'objets en classes caractérisées par leur similarité. Chaque classe représente un groupe d'objets partageant des similitudes internes et se distinguant des objets des autres classes. Ainsi, le clustering non supervisé repose sur la notion essentielle de similarité (ou de distance) . Il existe de multiples approches de clustering non supervisé qui utilisent différents algorithmes sous-jacents pour regrouper les points de données en fonction de leur similarité. Une approche simple du clustering non supervisé est le k-means. Ici, dans sa version de base le nombre de classes à identifier est prédéfini par un paramètre k fixé à l'avance. Chaque classe est représentée par un centre de classe, qui est un point de données artificielles représentant la valeur moyenne (ou médiane) de tous les points attribués à cette classe.

2.3.2 La réduction de la dimensionnalité

La réduction de la dimensionnalité est une technique qui permet de transformer des données de haute dimension en un espace de dimension inférieure, tout en conservant les informations importantes. Elle simplifie les données et améliore la performance des modèles d'apprentissage automatique. De nombreuses méthodes de réduction de dimensionnalité se concentrent sur la recherche de représentations de faible dimension de modèles de haute dimension, appelées variables latentes, représentations latentes ou plongements latents. Cette méthode utilise une transformation orthogonale pour changer les variables qui sont liées entre elles en un nouvel ensemble de variables qui n'ont plus de liens, et on les appelle les composantes principales. La réduction de dimensionnalité peut être utilisée pour diverses tâches, par exemple la visualisation, le prétraitement pour les méthodes de reconnaissance de formes ou pour les algorithmes symboliques. Pour permettre la compréhension et l'interprétation humaine des données de grande dimension, la réduction aux espaces a 2 et 3 dimensions est une tâche importante.

2.4 Autres méthodes d'apprentissage automatique

Il existe d'autres types d'apprentissage automatique tels que l'apprentissage semi-supervisé et l'apprentissage par renforcement. L'apprentissage semi-supervisé (SSL) se positionne entre l'apprentissage supervisé et non supervisé en traitant des ensembles de données comprenant à la fois des données étiquetées et non étiquetées. Contrairement à l'apprentissage supervisé, où toutes les données sont étiquetées, et a l'apprentissage non supervisé, où aucune étiquette n'est fournie, l'apprentissage semi-supervisé utilise partiellement des données étiquetées pour former le modèle tout en exploitant les données non étiquetées. Cette approche équilibrée est particulièrement adaptée aux situations où toutes les caractéristiques sont présentes, mais certaines n'ont pas de cibles associées. Ces circonstances se pressentent fréquemment lorsque l'étiquetage d'images est chronophage ou devient prohibitif. L'apprentissage semi-supervisé est couramment appliqué aux images médicales, où un médecin peut annoter un petit sous-ensemble d'images pour former le modèle. Ce dernier est ensuite utilisé pour classer le reste des images non étiquetées de

l’ensemble de données. L’ensemble de données étiqueté résultant est ensuite utilisé pour entraîner un modèle opérationnel qui devrait théoriquement surpasser les modèles non supervisés. En ce qui concerne l’apprentissage par renforcement, l’objectif est de développer un système capable d’apprendre en interagissant avec son environnement, de manière similaire au conditionnement opérant. Dans ce type d’apprentissage, le comportement de l’algorithme est façonné par une séquence de récompenses et de pénalités, dépendant de la précision de ses décisions par rapport à un objectif défini par le chercheur. Contrairement à l’apprentissage supervisé, où l’algorithme utilise des exemples pour modéliser le comportement, l’apprentissage par renforcement permet à l’algorithme d’explorer librement, c’est-à-dire par essais et erreurs, afin de déterminer quelles actions maximisent les récompenses et minimisent les pénalités. L’apprentissage par renforcement représente l’un des domaines les plus prometteurs de l’apprentissage automatique dans de nombreuses disciplines.

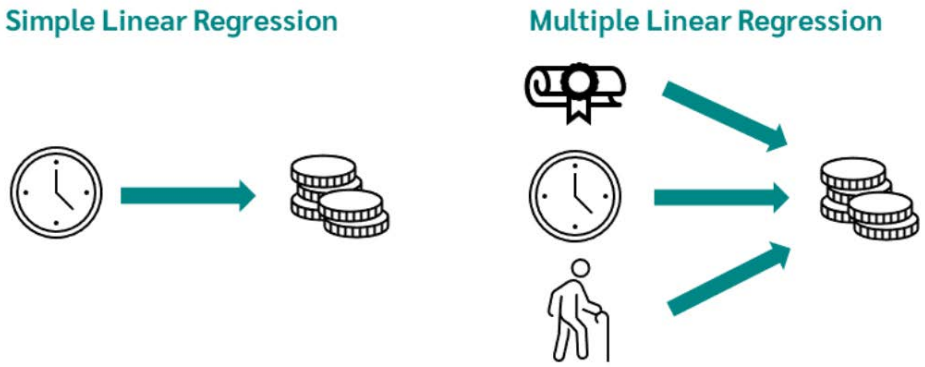
2.5 Quelques algorithmes de machine learning supervisé

2.5.1 La régression linéaire

En statistiques, un modèle de régression linéaire est un modèle de régression qui cherche à établir une relation linéaire entre une variable, dite expliquée, et une ou plusieurs variables, dites explicatives.

On parle aussi de modèle linéaire ou de modèle de régression linéaire.

L'analyse de régression linéaire est utilisée pour créer un modèle qui décrit la relation entre une variable dépendante et une ou plusieurs variables indépendantes. Selon qu'il y a une ou plusieurs variables indépendantes, on distingue l'analyse de régression linéaire simple et l'analyse de régression linéaire multiple **[12].**



[13]

Figure 2-5 : La différence entre la régression linéaire simple et une régression linaire multiple.

Dans le cas d'une régression linéaire simple, l'objectif est d'examiner l'influence d'une variable indépendante sur une variable dépendante. Dans le second cas, une régression linéaire multiple, on analyse l'influence de plusieurs variables indépendantes sur une variable dépendante.

Les équations nécessaires au calcul d'une régression simple et d'une régression multiple sont obtenues :

Simple Linear Regression

$$\hat{y} = b \cdot x + a$$

Multiple Linear Regression

$$\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a$$

Figure 2-6 : Les différentes équations de la régression linéaire simple et la régression linaire multiple.

Les coefficients peuvent être interprétés de manière similaire à l'équation de régression linéaire. Si toutes les variables indépendantes sont égales à 0, la valeur résultante est a.

Si une variable indépendante change d'une unité, le coefficient associé indique de combien la variable dépendante change. Ainsi, si la variable indépendante x_i augmente d'une unité, la variable dépendante y augmente de b_i.

Définition des coefficients de régression :

- ***a*** : le point d'intersection avec l'axe des ordonnées (y)
- ***b*** : la pente de la droite

$\hat{y}$ est l'estimation respective de la valeur y. Cela signifie que pour chaque valeur x, la valeur y correspondante est estimée. Dans notre exemple, cela signifie que la taille des personnes est utilisée pour estimer leur poids.

Le *coefficient* **de régression b** peut maintenant avoir différents signes, qui peuvent être interprétés comme suit :

- $b > 0$: il existe une corrélation positive entre x et y (plus x est grand, plus y est grand)
- $b < 0$: il existe une corrélation négative entre x et y (plus x est grand, plus y est petit)
- $b = 0$: il n'y a pas de corrélation entre x et y.

2.5.2 SVM linéaire

Un SVM linéaire **(Support Vector Machine linéaire)** est un algorithme d'apprentissage supervisé utilisé pour la classification et la régression. Il est particulièrement efficace pour séparer des données linéairement séparables en trouvant l'hyperplan optimal qui maximise la marge entre les classes.

Le SVM cherche un hyperplan sous la forme :

$$f(x) = \langle w, x \rangle + b \quad [14]$$

Où :

- w est le vecteur des poids,
- b est le biais.

L'algorithme maximise la marge entre les vecteurs de support, c'est-à-dire les points les plus proches de l'hyperplan.

Il minimise également une fonction de coût pour éviter le surajustement, en utilisant un paramètre de régularisation C.

L'optimisation repose sur la minimisation de l'expression suivante :

$$\frac{1}{2}\left|\left|w\right|\right|^2 + C\sum_i \xi_i \quad [15]$$

Où :

- ξ_i sont des variables de relaxation permettant de gérer les erreurs dans le cas où les données ne sont pas parfaitement séparables.

2.5.3 SVM radial

Le SVM (Support Vector Machine) avec un noyau Radial Basis Function (RBF) est un puissant algorithme d'apprentissage automatique utilisé pour les tâches de classification et de régression. Le noyau RBF, également connu sous le nom de noyau gaussien, est un choix populaire lorsqu'il s'agit de données séparables de manière non linéaire.

Le noyau RBF mappe les caractéristiques d'entrée dans un espace de dimension supérieure où une limite de décision linéaire peut séparer efficacement les classes. Il est particulièrement utile lorsque la relation entre les caractéristiques et les étiquettes est complexe et non linéaire.

Le noyau RBF est défini comme suit (Représentation mathématique) :

$$K(x, x') = \exp(-\gamma|x - x'|^2) \quad [16]$$

Où:

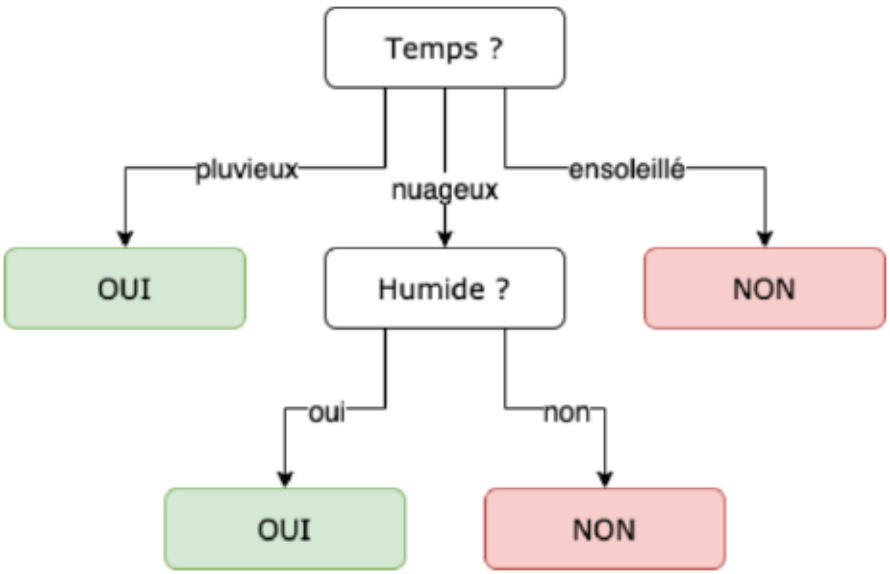
- x, x′ sont des vecteurs de caractéristiques
- γ (gamma) est un hyperparamètre qui contrôle l'influence d'un seul exemple d'entraînement. Des valeurs plus élevées conduisent à des modèles plus flexibles, tandis que des valeurs plus faibles créent des limites de décision plus fluides.

2.5.4 La forêt aléatoire

La forêt aléatoire est un algorithme d'apprentissage automatique basé sur un ensemble d'arbres de décision. Il est utilisé pour **les tâches de classification et de régression** et offre une excellente robustesse face au surajustement et aux données bruitées.

La forêt aléatoire est l'une de ces méthodes basée sur l'arbre de décision C A R T et la méthode de bagging. Le bagging consiste à construire plusieurs arbres de décision de manière indépendante, puis a agréger les prédictions en moyennant pour réduire la variance de prédiction. Contrairement à l'approche séquentielle du boosting, ou les poids des résultats précédents influent sur les poids actuels, le bagging construit les arbres de manière parallèle, exploitant ainsi efficacement les capacités des ordinateurs modernes.

La foret aléatoire est un ensemble d'arbres C A R T , et les randomisations interviennent dans deux aspects de l'algorithme. Dans la méthode de la foret aléatoires, on sélectionne de manière aléatoire l'ensemble d'entraînement $T_b(b = 1, \ldots, B)$ à partir de l'ensemble d'entraînement total T avec remplacement (c'est-à-dire, échantillonnage Bootstrap) pour entraîner chaque arbre C A RT . Les données laissées de cote lors de ce processus de sélection aléatoire sont appelées échantillons "out-of-bag". Lors de la construction de chaque arbre C A R T , la méthode de la foret aléatoire sélectionne aléatoirement M caractéristiques ou variables d'entrée parmi les P caractéristiques ($M < P$). La division optimale pour chaque arbre C A R T est calculée en fonction de T_b et des P caractéristiques sélectionnées.



[17]

Figure 2-7 : Un exemple d'arbre de décision.

2.5.5 Le lasso

LASSO **(Least Absolute Shrinkage and Selection Operator),** en statistiques et apprentissage automatique, est une méthode de régression qui permet à la fois :

- de sélectionner des variables,
- et de régulariser un modèle pour éviter le sur-apprentissage (overfitting).

La fonction de coût du LASSO ajoute une pénalité basée sur la **valeur absolue** des coefficients :

$$L_1 = \sum_{i=1}^{n} \left(y_i - \mu - x_i^T \beta\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \qquad [18]$$

Où λ contrôle la force de la régularisation.

La formulation du Lasso (Least Absolute Shrinkage and Selection Operator) remplace la pénalité L2 par une pénalité L1. Plus précisément, dans la régression Lasso, la fonction de coût est modifiée en ajoutant une pénalité proportionnelle à la somme des valeurs absolues des coefficients, ce qui favorise la réduction de certains coefficients à zéro. Cette

caractéristique permet de réaliser une sélection de variables, car les variables ayant des coefficients nuls sont exclues du modèle.

L'avantage principal du Lasso est donc sa capacité à produire des modèles plus simples et plus interprétables, en mettant en évidence uniquement les variables les plus pertinentes tout en éliminant celles qui sont moins significatives. Cette "parcimonie" est particulièrement précieuse pour l'interprétation des résultats, car elle permet de se concentrer sur les variables clés et d'éviter la surcharge d'informations, ce qui est souvent un problème dans les modèles avec un grand nombre de variables.

Cependant, le Lasso présente une limitation en imposant une certaine quantité de parcimonie. Il peut y avoir au plus p coefficients β_j non nuls. Cette limitation peut devenir significative lorsque le nombre de prédicteurs p dépasse le nombre d'observations n. De plus, il est souvent évoqué que la parcimonie induite par la pénalité L1 pourrait entrainer une moindre précision par rapport `a une pénalité L2. Dans les cas o ù plusieurs caractéristiques sont corrélées et ont des effets importants sur la réponse le Lasso a tendance à annuler certaines de ces caractéristiques, voire toutes sauf une. En revanche, la régression Ridge n'effectue pas une telle sélection, mais plutôt tend à ≪partager≫ la valeur des coefficients entre les prédicteurs corrélés.

2.5.6 La régression crête (en anglais: Ridge)

La régression Ridge (RR) et la régression Lasso (RL) sont des variantes régularisées de la régression des moindres carres qui appliquent respectivement des pénalités de type L2 et L1 sur le vecteur de coefficients. Considérons le modèle de régression linéaire multiple d´défini par $Y = X\beta + \varepsilon$.

Dans le contexte de la régression Ridge, l'objectif est de minimiser par rapport au vecteur de coefficients β, un critère formulé comme suit:

$$L2 = \sum_{i=1}^{N} \left(y_i - \mu - x_i^T \beta\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \qquad [\mathbf{19}]$$

Ou n représente le nombre d'observations, p est la dimension du vecteur de coefficient β, y_i est la valeur observée de la variable dépendante pour la i -ème observation, x_i est le vecteur des prédicteurs pour la i -ème observation, μ est l'intercept, β_j est le j-ème coefficient, et λ est le paramètre de régularisation de Ridge qui varie dans l'intervalle [0, ∞].

À mesure que la valeur de λ évolue, la solution β(λ) décrit un parcours dans l'espace des coefficients R^p. L'incorporation de la pénalité $\lambda \sum_{j=1}^{P} \beta_{2j}$ a pour effet de réduire la variance de l'estimation β(λ), mais cela peut également introduire un certain biais. Il est important de noter que le terme intercept μ ne figure pas dans la partie de la pénalité quadratique.

En introduisant $\varepsilon_i = y_i - \mu - x_i^T \beta$, la formulation de la régression Ridge peut également être interprétée comme la minimisation de la somme des carrés des résidus $|\epsilon|_2^2$ tout en imposant une contrainte de borne supérieure sur la norme L2 du vecteur de coefficients$|\beta|_2$. Cette contrainte conduit au même chemin dans l'espace β(λ), bien que chaque point le long de ce chemin puisse correspondre à une valeur différente de λ.

2.6 Rappel de quelques travaux scientifiques sur le cancer du sein

Il s'agit de travaux de recherche réalisés dans un domaine biomédical lié à la classification des cancers, en particulier du cancer du sein en utilisant des algorithmes d'apprentissage automatique (supervised Learning).

La Recherche sur l'algorithme de régression logistique des données de diagnostic du cancer du sein par apprentissage automatique a utilisé un algorithme de régression logistique pour

classer l'ensemble de données de patientes atteintes d'un cancer du sein, l'ensemble de données a été obtenu du référentiel **NIH** (National Library of Medicine). L'auteur a d'abord utilisé les 100 caractéristiques de l'ensemble de données pour entraîner le modèle et a finalement obtenu une précision de **65 %**. Ensuite, l'auteur, à l'aide d'une technique de sélection de caractéristiques, a extrait 46 caractéristiques principales des 100 les plus corrélés, à savoir la texture maximale, le périmètre maximal pour atteindre une précision de **87.89%,** ce qui est une amélioration par rapport au résultat obtenu à partir des 100 caractéristiques.

Les données de la mammographie ont été utilisées par le modèle de régression logistique pour prédire le facteur de risque de l'histoire du patient, la prédiction de la régression logistique est utilisée pour vérifier le pronostic des médecins et sont également utilisés pour corriger les prédictions incorrectes. Les travaux des auteurs peuvent aider les radiologues à diagnostiquer correctement le cancer du sein en utilisant la mammographie et en se référant aux antécédents de la patiente **[20].**

En utilisant Naïve Bayesian comme classificateur sur l'ensemble de données du Wisconsin de 10 caractéristiques l'auteur a essayé d'estimer le succès et l'erreur de l'algorithme de classification et de prédiction lorsque les données sont choisies au hasard **[21].**

Le modèle Naïve Bayes a montré un taux de réussite approximatif de **85 %** à **95 %** et un taux d'erreur de 10 à 15 % pour la classification et la prédiction. Dans leurs travaux de recherche, ont utilisé deux modèles d'apprentissage automatique : Naïve Bayesian et K Nearest voisin pour classer l'ensemble de données d'origine sur le cancer du sein du référentiel UCI Machine Learning. Leur objectif était de proposer lequel des deux est le plus efficace. En utilisant le même ensemble de données, ils y ont appliqué les différents algorithmes et en utilisant la validation croisée comme mesure de performance. Le résultat a montré que KNN avec **97,51%** pour la précision est légèrement meilleur que NB avec **96,19%.** Cependant, les auteurs ont suggéré qu'étant donné un ensemble de données plus important, le NB sera probablement plus performant parce que KNN sera affecté par sa complexité temporelle **[22]. Bazazeh & Shubair, (2016)** ont réalisé une étude comparative de trois algorithmes d'apprentissage automatique populaires pour la classification du cancer du sein : Support Vector Machine, Random Forest et Bayesian Network. Ils ont également utilisé l'ensemble de données original sur le cancer du sein du Wisconsin de l'UCI Machine Learning Repository **[23].**

Les auteurs ont utilisé la technique de validation croisée K- fold comme mesure de validation pour les classificateurs avec k = 10. Les paramètres utilisés pour leur comparaison étaient l'exactitude, la précision, le rappel et l'AUC ROC et après avoir effectué leur simulation sur l'ensemble de données avec les trois classificateurs, leur Le résultat montre que SVM a les performances les plus élevées en termes d'exactitude, de précision et de spécificité. Cependant, ils ont déclaré qu'en termes de classification correcte des tumeurs, le RF avait la probabilité la plus élevée.

En outre, **(Gupta & Gupta, 2018)** a effectué une analyse comparative de trois techniques d'apprentissage automatique largement utilisées, à savoir : le perceptron multicouche (MLP), l'arbre de décision (C4.5), la machine à vecteurs de support (SVM), le voisinage le plus proche (KNN) réalisée sur un ensemble de données sur le cancer du sein du Wisconsin pour prédire la récidive du cancer du sein. L'objectif principal de leur travail était d'obtenir le meilleur classificateur des quatre en termes d'exactitude, de précision, de rappel et de R2. Dans leur travail, ils ont conclu que la MLP était plus performante que d'autres techniques, et en plus, lorsque la métrique de validation croisée 10 fois était utilisée dans la prédiction du cancer du sein, la MLP avait également de meilleures performances **[24]. Khourdifi, Y., & Bahaj, M.** (2019) dans leurs travaux de recherche, ont appliqué quatre techniques d'apprentissage automatique, à savoir SVM, RF, Naïve Bayes et K-NN sur l'ensemble de données sur le cancer du sein du Wisconsin du référentiel d'apprentissage automatique de l'UCI. Les auteurs ont utilisé le logiciel Waikato Environment for Knowledge Analysis (Weka) pour la simulation de l'algorithme. Dans leurs résultats, SVM avait la performance globale en termes d'efficacité **[25].**

Enfin, Dans son mémoire de fin de cycle à l'UQTR, **Cheikh (2024**) a appliqué diverses techniques d'apprentissage automatique, notamment le SVM (linéaire et non linéaire), la régression linéaire, Lasso, Ridge ainsi que la forêt aléatoire, sur les mêmes jeux de données. En adoptant une approche similaire, il a obtenu une performance de 88 % en utilisant un ensemble de 94 variables. Cette performance a été rendue possible grâce à l'utilisation de la méthode de sélection de caractéristiques basée sur le coefficient de corrélation de Pearson, combinée à un modèle de forêt aléatoire **[26].**

2. 7 Conclusion

L'apprentissage automatique est un paradigme important et largement utilisé pour de nombreux problèmes. Dans ce chapitre, nous avons passé en revue les techniques de ML classiques et avancées. En particulier, nous avons présenté les déférents modèles d'apprentissage supervisé. La technologie ML est une bonne solution à de nombreux problèmes médicaux (tels que la détection du cancer du sein). Cependant, nous sommes encore loin de résoudre ce problème.

CHAPITRE 3 : ANALYSE STATISTIQUE DES DONNÉES

Introduction

Dans ce chapitre, la première partie est consacrée à la description des bases de données utilisées. Ensuite, nous présentons le cadre méthodologique mis en œuvre, ainsi que les résultats obtenus, en détaillant les différentes étapes réalisées à l'aide du langage de programmation choisi. Cela inclut les phases de prétraitement, d'apprentissage et de validation des modèles de prédiction. Des captures d'écran des résultats viennent illustrer et appuyer la démarche proposée.

3.1 Architecture du système

L'Architecture de notre système de prédiction pour la survie du cancer du sein à l'aide des Algorithmes d'Apprentissage Machine est la suivante :

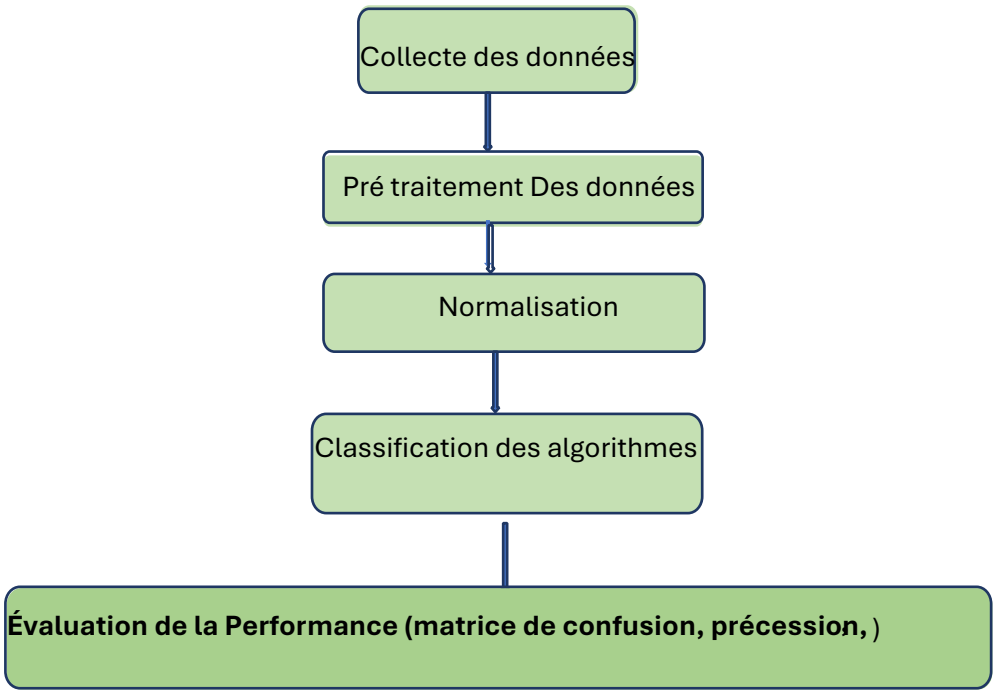


Figure 3-1 : Architecture de système pour l'apprentissage machine.

Le cadre proposé comprend les modules suivants : la collecte de données, l'étape de prétraitement qui implique le traitement des données manquantes, la formation et le test des modèles d'apprentissage automatique et enfin, l'analyse et la comparaison des performances. La **figure 3.1** illustre le cadre proposé, les données sont collectées à partir du référentiel d'apprentissage automatique en ligne provenant de **NIH** (National Library of Medicine) **[27]**. Les données collectées seront prétraitées, le prétraitement est effectué de manière à gérer les valeurs manquantes dans les données et une technique de mise à l'échelle des caractéristiques est utilisée pour normaliser les données si nécessaire. Nous avons deux jeux de données; un ensemble pour l'apprentissage et l'autre pour le test. L'ensemble de jeux de données pour l'apprentissage est utilisé pour former les modèles de prédiction, tandis que l'ensemble de test est utilisé à des fins de validation. À l'aide de ces mesures de performance, qui sont la précision, le rappel, Mean Absolute Error (MAE), Mean Squared Error (MSE) et le score f1, les modèles de prédiction sont évalués et comparés.

3.2 Les outils matériels et logiciels

3.2.1 Partie logicielle

Le choix d'un environnement de programmation adéquat est crucial pour la progression de notre projet. Cette décision est influencée par divers éléments, tels que la capacité de montage, la simplicité d'usage, l'accessibilité des fonctionnalités, l'interaction avec d'autres environnements, et bien plus encore.

Nous avons opté pour *PYTHON* qui est optimisé pour résoudre les problèmes scientifiques et techniques. Le langage PYTHON, basé sur les matrices un moyen naturel pour exprimer les mathématiques computationnelles. PYTHON est l'un des langages les plus accessible et les plus productif conçu pour les ingénieurs et les scientifiques. Il est doté de plusieurs éditeurs (ex: Jupyter Notebook) permettant d'exécuter des séquences de commandes encapsulées dans des fonctions.

Grâce à sa vaste bibliothèque de boîtes à outils prédéfinies, nous pouvons commencer directement par les algorithmes essentiels à notre domaine.

Le code PYTHON peut être intégré à d'autres langages, ce qui nous permet de déployer des algorithmes et des applications au sein de systèmes Web, d'entreprise et de production. À cela s'ajoute les graphiques intégrés permettant de visualiser facilement les données afin d'en dégager des informations. Il offre aussi des applications dédiées à l'ajustement de courbes, la classification de données, l'analyse de signaux et bien d'autres tâches spécialisées.

3.2.2 Partie matériel

Notre système a été conçu sur une machine avec pour :

Caractéristiques	***Information***
Processeur	*AMD Ryzen 5 4000 Series with Radeon Graphics 3.30 GHz*
Mémoire vive installée	*16,0 Go (15,2 Go utilisable)*
Type du système	*Système d'exploitation 64 bits, processeur x64*
Édition	*Windows 11 Professionnel*
Version	*23H2*
Version du système d'exploitation	*22631.3737*

Tableau 3.1 – Caractéristiques du système d'exploitation.

Notre environnement de programmation n'a pas besoin de GPU (Unité de Traitement Graphique) ou de TPU (Unité de Traitement du Tenseur) pour fonctionner. Cependant, ces accélérateurs matériels avec au moins 4 Go de RAM sont recommandés pour exécuter efficacement des tâches d'apprentissage automatique sur de grands ensembles de données.

3.3 Description de la base de données

Le jeu de données utilisé dans cette étude regroupe des informations génétiques et cliniques relatives à un ensemble de patientes atteintes de cancer du sein. Parmi les variables disponibles figure notamment la durée de vie avant le décès, exprimée en nombre de jours et désignée par l'intitulé *days_to_death*.

Ce mémoire a pour objectif de développer un modèle capable de prédire la durée de survie des patientes à partir de ces données. L'approche adoptée repose sur l'utilisation de deux ensembles distincts : un ensemble de découverte (discovery_data), intégrant à la fois des données cliniques et génétiques, et un ensemble de validation (validation_data), également composé de données cliniques et génétiques, destiné à valider voire évaluer la performance du modèle.

L'ensemble de découverte clinique comprend un total de 509 patientes atteintes de cancer du sein. Pour chacune d'elles, plusieurs variables cliniques sont prises en compte, notamment l'âge au moment du diagnostic pathologique, le stade de la tumeur, les traitements reçus, l'état vital, la taille de la tumeur ainsi que le temps observé jusqu'au décès.

L'ensemble de découverte génétique correspond à celui de l'ensemble de découverte clinique, contenant le même nombre de patientes, soit 442. Il inclut les niveaux d'expression génique mesurés pour un total de 24 925 gènes, permettant une analyse approfondie des profils moléculaires associés à la survie.

Les ensembles de validation, à la fois clinique et génétique, sont quant à eux utilisés pour évaluer la performance du modèle de prédiction développé. Ces données sont totalement indépendantes de l'ensemble de découverte.

Le groupe de validation est constitué de patientes indépendantes de celles utilisées dans l'ensemble de découverte. La recherche de gènes pouvant être utilisés pour prédire le pronostic, en particulier la durée de survie chez les patientes atteintes de cancer du sein, revêt une importance majeure.

Une telle identification permettrait d'affiner les options thérapeutiques et de faire progresser notre compréhension des mécanismes moléculaires impliqués dans la progression tumorale.

Cependant, l'hétérogénéité importante des échantillons de patientes constitue un obstacle majeur à l'identification de gènes véritablement pronostiques, et complique la prédiction précise des résultats cliniques. Cette diversité biologique rend les modèles plus sensibles aux variations individuelles, limitant parfois leur capacité de généralisation.

Les données initialement collectées étaient structurées sous forme de dictionnaire. La première étape du prétraitement a consisté à transformer ces données en un *DataFrame*, où les identifiants des patientes apparaissaient en lignes, et les gènes en colonnes. Une opération de transposition a ensuite été effectuée afin de corriger cette structure, aboutissant à une représentation conforme aux standards d'analyse, avec les gènes en lignes et les patientes en colonnes.

	473	645218	494470	...	54862	57549	149647	Days_to_death
MB 0002	3,42466	2.68855	2,90461	...	2,87343	3.06007	2.62570	2002
MB 0008	3.36815	2.67866	2,91721	...	2,86258	2.79879	2.66465	453
MB 0010	3.31097	2.64375	2,90304	...	2,85483	3.06626	2.64078	1729
MB 0035	3.05326	2.63734	2,64231	...	2,91272	3.00762	2.56304	325
MB 0036	3.18358	2.63330	2,74730	...	2,83853	3.03667	2.71863	1585
MB 0050	3.28523	2.69370	3,15683	...	2,95563	3.16382	2.61689	1927
MB 0059	3.26687	2.63087	3,27462	...	2,80160	2.80758	2.67339	1195
MB 0060	3.14576	2.69921	3,00717	...	2,94174	2.95289	2.61004	606
MB 0066	3.28141	2.67614	2,89571	...	2,85755	3.09616	2.65825	2083
MB 0101	3.29434	2.69136	3,13095	...	2,80261	3.08000	2.69208	3628

Table 3.2 – Données de patients atteints de cancer.

L'objectif de ces travaux est multiple :
– Identifier de nouveaux biomarqueurs pour affiner le diagnostic.
– Personnaliser les traitements en fonction du profil génétique des patientes.
– Améliorer la classification du cancer du sein pour une prise en charge plus ciblée.
– Optimiser les stratégies thérapeutiques en combinant les données cliniques et génétiques.

Les bases de données utilisées pour soutenir cette recherche sont :
NCBI (National Center for Biotechnology Information),

TCGA (The Cancer Genome Atlas),
METABRIC (Molecular Taxonomy of Breast Cancer International Consortium),
cBioPortal,
ICGC (International Cancer Genome Consortium).

3.4 Prétraitement de la base des données

Les données à disposition des entreprises sont souvent désordonnées et de mauvaise qualité, ce qui représente un frein dans le processus car il est nécessaire de passer beaucoup de temps à améliorer ces données avant de passer à l'analyse.

Le prétraitement ou la préparation de données est une étape cruciale dans le processus d'exploration de données. C'est une étape qui précède celui de l'analyse de données. Elle est constituée de plusieurs tâches comme la collecte de données, le nettoyage de données, l'enrichissement de données ou encore la fusion de données **[28].**

Au cours de la préparation des données, les données dites « brutes » sont soumises à différents traitements afin de les rendre exploitables pour l'étape d'exploration de données, au cours de laquelle le but sera d'extraire des connaissances à partir des données via la construction de modèles.

La préparation des données une étape clé car la fiabilité de l'analyse des données dépend en très grande partie de la qualité des données.

Ce processus permet d'améliorer la qualité des données et de faciliter l'application de modèles d'analyse ou d'apprentissage automatique.

Tous d'abord nous avons importé les librairies nécessaires pour la préparation des données:

```
## import libraries

import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline

# Models  from  scikit-learn

from sklearn.model_selection import train_test_split, KFold, cross_val_score
from sklearn.ensemble import RandomForestRegressor
from sklearn.linear_model import LinearRegression,Lasso, Ridge
from sklearn.svm import LinearSVR, SVR,SVC


# Model  evalutaions

from lifelines.utils import concordance_index
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
from sklearn.metrics import accuracy_score, classification_report
```

Figure 3-2 : Représente le code python utilisé pour importer la librairie.

Les données téléchargées à partir du référentiel d'apprentissage automatique sont dans mon ordinateur : " C:\Users\SORO\Desktop\Sample_project".

La programmation Python a une bibliothèque connue sous le nom de Pandas, qui peut être utilisée pour ouvrir et lire des fichiers à valeurs séparées par des virgules (CVS).

```
discovery_data.isnull().sum()
```

```
473              0
645218           0
494470           0
641719           0
51533            0
                ..
114788           0
54862            0
57549            0
149647           0
days_to_death    0
Length: 24925, dtype: int64
```

Figure 3-3 : Afficher les données nulles sur la base (découverte).

L'analyse réalisée à l'aide de la fonction discovery_data. isnull().sum() indique qu'il **n'y a aucune valeur manquante** dans le jeu de données, toutes les colonnes renvoyant une valeur égale à **0**. En effet, cette fonction retourne le nombre de valeurs nulles par colonne : une valeur de **0** signifie qu'aucune donnée n'est absente, tandis qu'une valeur supérieure à 0 indiquerait la présence de données manquantes.

3.5 Analyse exploratoire des données

La durée de survie moyenne pour ce groupe de patients atteints de cancer du sein est d'environ 2698.247 jours, soit environ **7 ans et 5 mois**, ce qui donne une idée générale du pronostic pour l'ensemble de la cohorte. Toutefois, un écart type de 1786.497 jours révèle une forte variabilité des durées de survie, reflétant la diversité des parcours cliniques. Cette variabilité peut être attribuée à plusieurs facteurs, notamment le stade du cancer au moment du diagnostic, le type de traitement reçu, ainsi que des caractéristiques individuelles telles que l'âge ou les comorbidités.

L'intervalle de survie, allant de 43 à 8220 jours, met en évidence l'ampleur de cette hétérogénéité : certains patients décèdent très rapidement après le diagnostic, tandis que d'autres survivent pendant plus de 22 ans, illustrant la complexité biologique du cancer du sein.

La médiane, située à 2163 jours (**5 ans et 11 mois**), indique que la moitié des patients vivent au-delà de cette durée, suggérant qu'un grand nombre d'entre eux bénéficient d'une espérance de vie prolongée, malgré la gravité potentielle du diagnostic.

Enfin, les quartiles permettent de mieux comprendre la distribution des durées de survie :

- Le premier quartile **(Q1)** est à 1235 jours (soit environ **3 ans et 5 mois**),
- Le troisième quartile **(Q3)** à 4131 jours (soit environ **11 ans et 4 mois**.), ce qui signifie que **50 %** des patients ont une durée de survie comprise entre ces deux valeurs. Cette répartition souligne encore une fois la variabilité interindividuelle face à la maladie.

```
count     509.000000
mean     2698.247544
std      1786.497388
min        43.000000
25%      1235.000000
50%      2163.000000
75%      4131.000000
max      8220.000000
Name: days_to_death, dtype: float64
Minimum : 43.0 jours
Maximum : 8220.0 jours
Moyenne : 2698.2475442043224 jours
Médiane : 2163.0 jours
Écart-type : 1786.4973884437113 jours
```

Figure 3.4 : Code Python des Statistiques sur la durée de survie (days_to_death) pour le cancer du sein.

Durée de survie (jours)	
Moyenne (jours)	2698
Écart type	1786
Minimum (jours)	43
Maximum (jours)	8220
Médiane (jours)	2163
Premier quartile (Q1)	1235
Troisième quartile (Q3)	4131

Table 3.3 – Statistiques sur la durée de survie (days_to_death) pour le cancer du sein.

L'histogramme de la durée de survie offre un aperçu détaillé de la distribution des données, montrant la variabilité, la tendance centrale et la diversité des expériences des patients atteints du cancer du sein dans notre échantillon. Dans la Figure 4.1, on observe que les patients se trouvant dans les intervalles [500, 2500] jours et [3500,5500] jours ont une durée de survie beaucoup plus longue que les patients situés dans les intervalles [0, 500] jours, [2500, 3000] jours et [5500, 8220] jours.

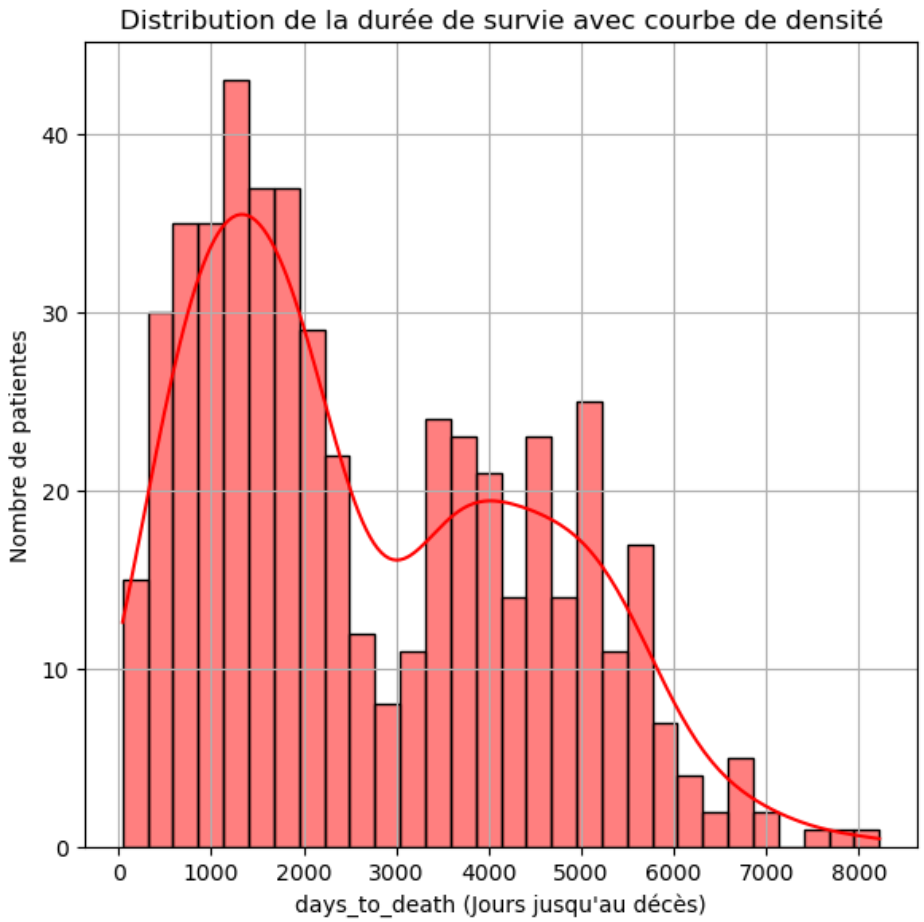


FIGURE 3.5 – Distribution de days_to_death

3.5.1 Modèle univarié : Association entre un gène et la survie

Le modèle de régression de Cox a été appliqué afin d'évaluer l'impact de l'expression des différents gènes sur la survie des patientes atteintes de cancer du sein. Comme expliqué dans le chapitre 2, cette méthode statistique permet de modéliser la durée de survie en prenant en compte l'influence de plusieurs variables explicatives. En l'occurrence, elle est utilisée ici pour identifier les gènes dont l'expression est significativement associée à la probabilité de survie.

Les résultats de cette analyse sont présentés dans la figure 3.6.

Out[65]:

	Gene	coef	HR (exp(coef))	p-value	significant
7	4170	1.609634	5.000983	0.000305	True
52	29945	-2.628583	0.072181	0.000819	True
86	6348	1.087577	2.967077	0.000875	True
81	80042	-1.906730	0.148565	0.000919	True
26	55016	1.911247	6.761514	0.001014	True
54	57654	-1.528616	0.216835	0.001124	True
49	9442	1.978514	7.231988	0.001252	True
80	100008589	1.155129	3.174434	0.001714	True
79	135114	1.965599	7.139189	0.001839	True
62	732007	1.450717	4.266172	0.001900	True
67	5127	2.780816	16.132179	0.002035	True
8	79647	1.472557	4.360370	0.002101	True
83	401238	1.981394	7.252845	0.002346	True
87	731878	2.316834	10.143508	0.002814	True
29	653158	-1.214829	0.296761	0.002969	True
23	7852	0.918941	2.506635	0.003260	True
93	5223	1.422266	4.146507	0.003305	True
19	158295	-1.082947	0.338596	0.004008	True
72	467	0.784159	2.190564	0.005159	True
16	8915	1.883559	6.576871	0.005633	True

Figure 3.6 - Résultats du modèle de régression de Cox pour les gènes les plus significatifs dans la survie.

Sur l'ensemble des 100 variables les plus corrélées retenues dans l'étude de l'association avec la survie, représentée par la variable ***days_to_death*** ont dénoté 63 *variables* significatives. Les 5 plus importantes sont les suivantes :

Gène 4170 :

Coefficient (coef) : Le coefficient représente la variation attendue dans le logarithme du taux de risque pour une unité d'augmentation de la variable 4170. Dans ce cas, un coefficient de 1,60 suggère une augmentation significative du risque.

Exp(coef) : L'exponentielle du coefficient est interprétée comme le facteur multiplicatif

Par lequel le taux de risque change pour une unité d'augmentation de la variable121260.

Ici, exp (1,60) équivaut à environ 5.000983. Cela signifie que, pour une unité d'augmentation dans la variable 4170, le risque est multiplié par environ 5.000983.

Valeur de p : La valeur de p (p-value) est très faible $\mathbf{3,05 \times 10^{-4}}$, indiquant une significativité statistique. Cela suggère que le coefficient n'est probablement pas nul, et la relation entre la variable 4170 et le résultat est statistiquement significative.

De manière analogue pour les variables :

Gène 29945 :

Une augmentation d'une unité de cette covariable entraîne une augmentation significative de 51,28 fois du risque de l'évènement. La faible valeur de p suggère une forte signification statistique, renforçant l'impact.

Gène 6348 :

Une augmentation d'une unité de cette covariable correspond à une augmentation

Considérable de 14,62 fois du risque de l'évènement. La valeur de p est assez faible,

Indiquant un soutien statistique significatif pour cet effet.

Gène 80042 :

Une augmentation d'une unité de cette covariable est associée à une augmentation de

3,65 fois du risque de l'évènement. La valeur de p est faible, signifiant une signification

Statistique, bien que moins forte que les précédentes.

Gène 55016 :

Une augmentation d'une unité de cette covariable entraîne une augmentation de 11,9 fois du risque de l'évènement. La valeur de p est modérément faible, soutenant la signification

statistique de cet effet. Ces interprétations mettent en évidence l'ampleur des changements des rapports de risques et la signification statistique associée.

La courbe de survie est un outil d'analyse de survie largement utilise pour visualiser la probabilité de survie au fil du temps. La courbe commence à 1 (100 % de probabilité de survie)

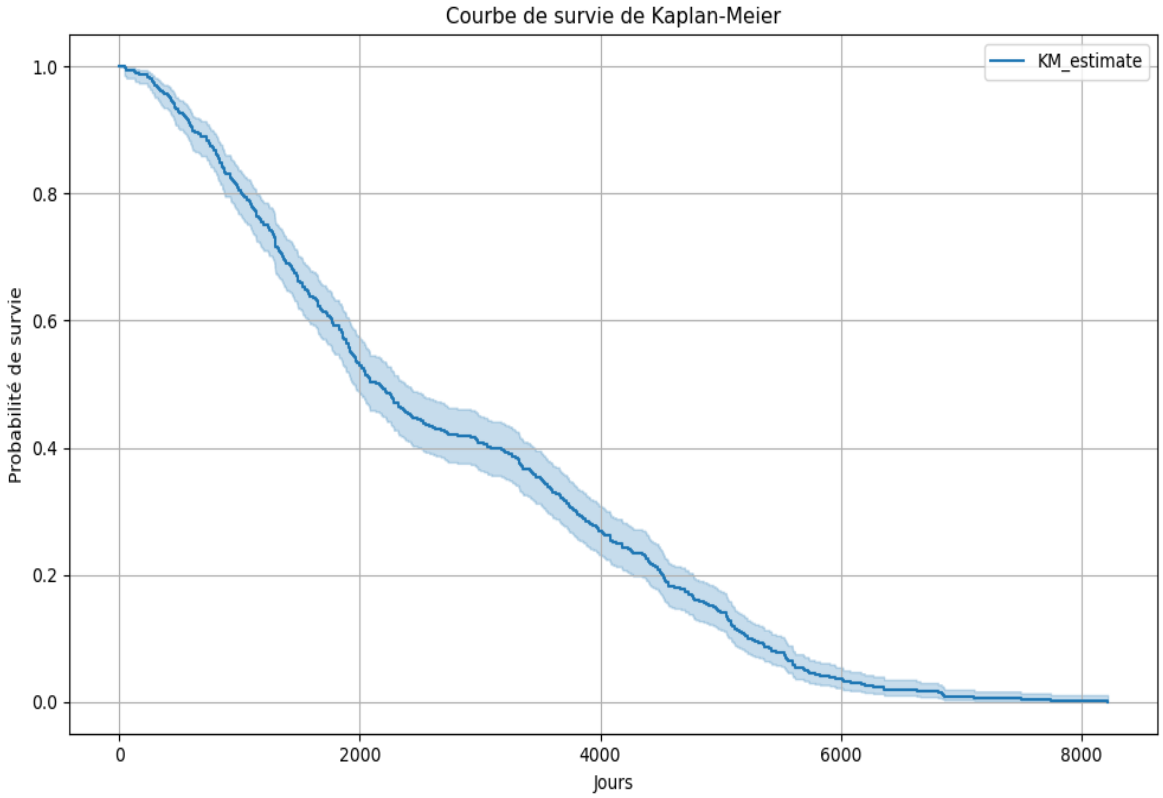


Figure 3.7 – Courbe de Survie.

A l'instant initial et diminue à mesure que le temps avance. Chaque descente de la courbe représente un évènement de décès ou de défaillance. Les paliers horizontaux entre les descentes indiquent des périodes ou aucun évènement n'a eu lieu. La forme de la courbe de survie peut fournir des informations précieuses sur les données de survie. Une courbe qui descend rapidement indique une diminution rapide de la probabilité de survie, ce qui peut signifier que l'évènement d'intérêt à une forte influence sur la survie. A l'inverse, une courbe qui descend lentement suggère une probabilité de survie relativement élevée au fil du temps.

Dans la figure 3.7, on observe que la courbe de survie descend rapidement, indiquant une diminution rapide de la probabilité de survie. Cela confirme les résultats trouvés sur la distribution de survie : plus le temps augmente, plus la probabilité de survie diminue.

3.6 Modèle multivarié

Nous avons différentes méthodes pour sélectionner les 100 meilleures variables parmi les 24925 gènes en relation avec la survie, notamment la corrélation Pearson, l'information mutuelle et la corrélation de Spearman. La sélection des caractéristiques joue un rôle essentiel dans l'amélioration des performances prédictives des modèles, et cette étude explore trois approches pour identifier les variables les plus pertinentes et ainsi affiner la qualité de la prédiction : la corrélation de Pearson, la corrélation de Pearson et l'information mutuelle.

3.6.1 Modèle base sur la corrélation et approches en apprentissage automatique

Pour la méthode de sélection de caractéristiques de Pearson, nous avons calculé la corrélation de Pearson en valeur absolue pour l'ensemble des 24925 gènes par rapport à la survie. Nous avons sélectionné les 100 meilleures caractéristiques qui sont les plus corrélées avec la survie. Ensuite, nous avons calculé les statistiques descriptives de la corrélation des variables sélectionnées. Les résultats obtenus sont représentés dans le tableau 3.4

Corrélation	
Minimum	0.2584
Premier quartile (Q1)	0.2638
Médiane	0.2725
Troisième quartile (Q3)	0.2861
Maximum	0.3236

TABLE 3.4 – Statistiques sur la Durée de Survie (days_to_death) pour le Cancer du Sein.

Les statistiques révèlent une gamme captivante de valeurs de corrélation, reflétant la diversité des liens potentiels entre les gènes et les résultats de survie. Parmi les valeurs présentées, on observe un éventail allant d'une corrélation minimale de 4.3 a une corrélation maximale de 8.220. Cette dispersion souligne la variabilité des interactions possibles entre les gènes et la survie, offrant ainsi un aperçu des relations subtiles et complexes au sein de l'ensemble de données. Les quartiles de corrélation fournissent également des repères essentiels pour saisir la distribution des valeurs. Le premier quartile, a 1.235, indique que 25% des gènes présentent une corrélation supérieure à ce seuil, tandis que la médiane, a 0.05, illustre la médiane des valeurs de corrélation, signalant ainsi le point central de cette distribution. Le troisième quartile, a 4.131, met en évidence la proportion de gènes ayant des corrélations encore plus marquées avec la survie. Ces chiffres suggèrent que les gènes peuvent avoir des degrés variables de corrélation avec la survie, allant des liens plus faibles aux associations plus prononcées. Il est crucial de souligner que ces mesures de corrélation offrent un aperçu statistique des relations potentielles et ne tiennent pas compte des interactions biologiques complexes qui peuvent sous-tendre ces observations. Nous avons appliqué de manière analogue la méthode de sélection de caractéristiques de Spearman et celle de l'information mutuelle.

En somme, ces résultats apportent un éclairage précieux sur la complexité des liens entre les gènes et la survie, incitant à des investigations plus approfondies pour comprendre les implications biologiques et cliniques de ces découvertes.

Nous avons évalué sept modèles de régression couramment utilisés dans la prédiction médicale en utilisant un jeu de données comprenant des informations sur des patients atteints de cancer du sein et le temps qui s'est écoulé avant leur décès. Les modèles étudiés sont tous détaillés dans le chapitre 3, qui sont les suivants : foret aléatoire, régression Ridge, régression Lasso, régression linéaire, régression élastique net, SVM non-linéaire et SVM cas linéaire. Pour chaque modèle, nous avons enregistré le nombre de variables utilisées et le coefficient de détermination R^2 pour évaluer les performances de prédiction.

3.7 Étude comparative

3.7.1 Méthode de sélection de caractéristiques de Pearson

Après avoir utilisé la méthode de sélection de Pearson pour sélectionner les 100 variables les plus importantes dans la survie, nous avons mis en place notre premier modèle de régression linéaire en utilisant une boucle pour évaluer les modèles avec différentes combinaisons de caractéristiques. À chaque itération de la boucle, un modèle de régression linéaire est entraîné sur l'ensemble d'entrainement avec validation croisée. Les prédictions sont ensuite faites sur l'ensemble de validation et les métriques d'évaluation telle que l'erreur absolue moyenne (MAE) et l'erreur quadratique moyenne (MSE) sont calculées.

Nous avons Également calculé la corrélation et l'indice de concordance entre les valeurs réelles et prédites. Enfin, nous avons tracé un graphique montrant la performance R^2 en fonction du nombre de variables, et nous avons utilisé cette performance dans cette étude pour évaluer les différents modèles. Nous avons répété le même processus pour tous les modèles dans les différentes méthodes de sélection de caractéristiques. Le code python utilisé pour le modèle de régression linéaire est représenté dans la figure 3.8.

La figure 3.8 représente la performance en fonction du nombre de variables. Le résultat trouvé est égal à 0.10095, ce qui indique généralement un ajustement modéré et bon, montrant la capacité du modèle de régression linéaire à fournir une meilleure explication de la variable cible days_to_death. Le nombre de variables associées est égal à 8 variables.

```
# Calculate R² and MAE on the test set
r2 = r2_score(y_test, predictions)
mae = mean_absolute_error(y_test, predictions)

r2_values[i - 1] = r2
mae_values[i - 1] = mae
```

```
r2
```

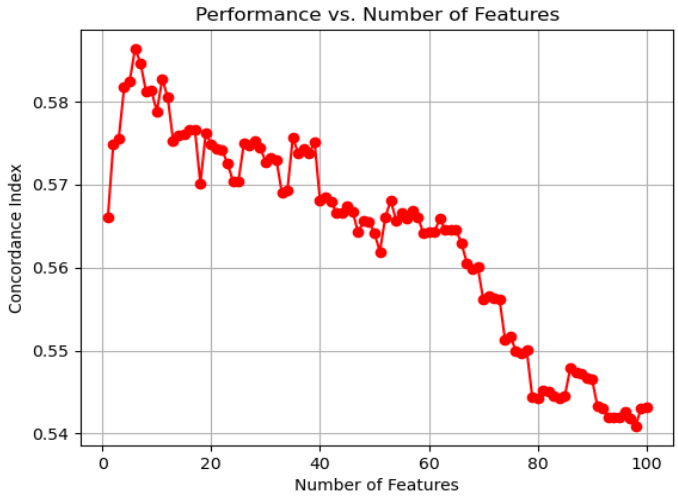
```
0.10095920504369715
```

```
mae
```

```
1440.2848039215692
```

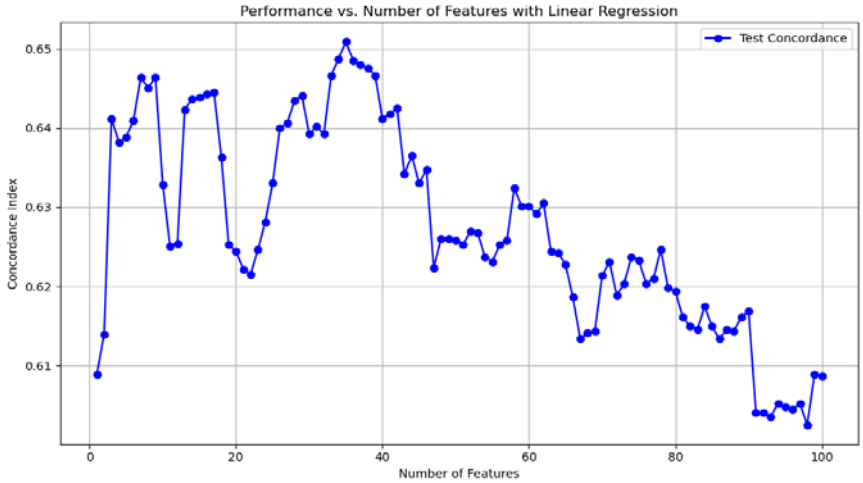
F I G U R E 3.8 – Performance du modèle de régression linéaire pour la méthode de sélection de caractéristiques de Pearson.

La figure 3.9 représente une combinaison de la corrélation entre les valeurs réelles et les valeurs prédites en fonction du nombre de variables dans le modèle de régression linéaire.



F I G U R E 3.9 – Corrélation entre les valeurs réelles et les valeurs prédites en fonction du nombre de variables.

La figure 3.10 montre une corrélation maximale de 0.65, ce qui indique un bon ajustement et une corrélation positive modérée entre les valeurs réelles et prédites. Le nombre maximal de variables associées est égal à 35 variables. Cela suggère que les valeurs prédites augmentent généralement avec les valeurs réelles et vice versa, mais cette relation est positive et forte.



F I G U R E 3.10 – indice de concordance entre les valeurs réelles et les valeurs prédites en fonction du nombre de variables.

La figure 3.10 représente l'indice de concordance entre les valeurs réelles et les valeurs prédites en fonction du nombre de variables dans le modèle de régression linéaire. La figure 4.5 montre un indice de concordance maximale de 0,6509, ce qui indique un accord substantiel entre les classements des valeurs réelles et prédites dans le modèle de régression linéaire.

Le nombre maximal de variables associées est de.

En d'autres termes, les positions relatives des observations dans les deux ensembles de données sont similaires. Cette valeur élevée de l'indice de concordance suggère que le modèle est capable de reproduire avec précision les ordres des valeurs réelles dans ses prédictions.

Par la suite, nous avons décidé de présenter les résultats sous forme de tableaux pour l'ensemble des modèles, dans le but de faciliter leur comparaison et d'en améliorer la lisibilité.

Modèles	Nombre de variables	R^2	MAE
Forêt aléatoire	88	0.0794	1538,96
Régression Ridge	89	0.0741	1523.3
Régression Lasso	34	0.0255	1581
Régression Linéaire	6	0.10095	1440.28
Régression Net Élastique	82	0.0770	1506.8
S V M non linéaire	2	0.0231	1508.6
S V M linéaire	41	0.0250	1507.8

T A B L E 3.5 – Tableau comparatif des performances des modèles de régression pour la méthode de sélection de caractéristiques de Pearson.

Les résultats de notre étude comparative montrent une variabilité significative dans les performances des différents modèles. Le modèle de Régression Linéaire se démarque en affichant la meilleure performance parmi tous les modèles évalués. Son coefficient de détermination R^2 élevé de 0,10095 suggère qu'il est capable d'expliquer une grande partie de la variance dans les données. De plus, sa MAE (Erreur Absolue Moyenne) relativement faible de 1440,28 indique que les prédictions du modèle sont proches des valeurs réelles. Ces résultats indiquent que le modèle de Régression Linéaire offre des prédictions précises et robustes. En revanche, les modèles de régression régularisée, tels que la Régression Ridge et la régression Lasso, présentent des performances similaires en termes de MAE et de R^2 légèrement distinct, bien que légèrement inférieures à celles du Foret aléatoire. Malgré cela, ils montrent un R^2 relativement faible qui suggère qu'ils ne parviennent pas à capturer la structure sous-jacente des données. Quant aux modèles de Régression Nette Élastique, SVM Radial et SVM Linéaire, ils affichent des performances plus faibles, caractérisées par des R^2 proches de zéro ou modérés et des MAE plus élevées. Cela suggère qu'ils ont du mal à prédire avec précision les valeurs de survie, ce qui peut être attribue à leur incapacité à capturer la complexité de la relation entre les caractéristiques et la variable cible.

Les résultats de notre étude comparative montrent une variabilité significative dans les performances des différents modèles. Le modèle de Régression Linéaire se démarque avec un R^2 modéré de 0,10095, indiquant une bonne capacite à prédire le nombre de jours avant le décès. En revanche, la Forêt aléatoire, la Régression Nette Élastique, et les SVR cas linéaire et non linéaires présentent des R^2 relativement faibles, suggérant une moindre capacite à expliquer la variance des données.

Modèles	Nombres de variables	Corrélation de Pearson
Foret Aléatoire	2	0,47107
Régression Ridge	82	0,43745
Régression Lasso	31	0,44409
Régression Linéaire	35	0,6509
Régression élastique Net	19	0,36034
S V M non linéaire	49	0,4017
S V M linéaire	19	0,3315

Table 3.6 – Corrélations de Pearson maximales entre les valeurs réelles et prédites avec les nombres de caractéristiques maximales associées pour les différents modèles.

Les résultats obtenus reflètent des variations intéressantes dans les performances des modèles en fonction de la corrélation de Pearson. Voici ce que nous avons observé pour chaque modèle : Régression Linéaire (Corrélation : 0,6509) : La corrélation positive suggère une cohérence entre les valeurs prédites et les valeurs réelles. Cela peut indiquer que le modèle Linear Regression capture efficacement les tendances et les variations des données. Random Forest (Corrélation : 0,47107) : Bien que largement inférieure à celle du Linear Regression, cette corrélation dénote toujours une relation significative entre les prédictions et les valeurs réelles. Le modèle Random Forest parvient à saisir certaines des tendances des données. Régression Lasso (Corrélation : 0,44409) : Avec une corrélation légère à celle du Random Forest, la régression Lasso montre également une corrélation positive. Cela indique une correspondance entre les prédictions du modèle et les valeurs réelles. Régression Ridge (Corrélation : 0,43745) : Bien que similaire et légèrement inférieure à celle de la Régression Lasso, la corrélation positive de la Régression Ridge montre que ce modèle parvient à capter

des tendances générales dans les données. SVM Non-linéaire (Corrélation : 0,4017) : Une corrélation inférieure suggère que ce modèle peut nécessiter des ajustements pour mieux saisir les subtilités des données et fournir des prédictions plus précises.

Modèles	Nombres de variables	Indice de concordance
Forêt Aléatoire	2	0,4654
Régression Ridge	97	0.3460
Régression Lasso	71	0.335
Régression Linéaire	8	0,5864
Régression élastique Net	97	0.3552
S V M non linéaire	41	0,4630
S V M Linéaire	7	0.3070

T A B L E 3.7 – Indice de concordance maximale entre les valeurs réelles et prédites avec les nombres de caractéristiques maximales associées pour les différents modèles.

Dans le domaine de la prédiction de la survie du cancer du sein, l'accord entre les valeurs prédites par les modèles et les valeurs réelles revêt une importance capitale pour la prise de décisions cliniques éclairées. Le modèle de la Regression Linéaire se distingue en affichant un indice de concordance élevé de 0,5864, ce qui suggère un accord solide entre les prédictions et les résultats observés, permettant ainsi de fournir des estimations précises de la survie des patients. Les modèles de Régression Ridge, Lasso et linéaire présentent des indices de concordance similaires, tous autour de 0,35, indiquant également un accord relativement faible entre les prédictions et les valeurs réelles dans le contexte de la survie du cancer du sein. Ces modèles offrent ainsi une estimation faible de la survie des patients, ne facilitant pas la prise de décisions médicales. En revanche, les indices de concordance des modèles de Régression Élastique Net, SVM non linéaire et SVM linéaire sont légèrement inférieurs, ce qui suggère un accord légèrement moins robuste entre les prédictions et les valeurs réelles pour ces modèles dans le cadre spécifique de la survie du cancer du sein. Bien que ces modèles puissent fournir des prédictions utiles, leur capacite à estimer précisément la survie des patients peut être légèrement limitée par rapport aux autres modèles.

3.7.2 La méthode de sélection de caractéristiques de Spearman

Le tableau 3.8 compare les performances des différents modèles de régression en utilisant la méthode de sélection de caractéristiques de Spearman.

Modèles	Nombres de variables	R^2	MAE
Forêt Aléatoire	24	0.1010	1440.284
Régression Ridge	96	-0.08695	1625.410
Régression Lasso	69	-0.02496	1581.024
Régression Linéaire	94	0.08131	1550,801
Régression élastique Net	70	-0.04208	1614.922
S V M non linéaire	49	-0.02263	1508.600
S V M linéaire	69	-1.3365	2151.087

T A B L E 3.8 – Tableau comparatif des performances des modèles de régression pour la méthode de sélection de caractéristiques de Spearman.

Les résultats de notre étude comparative révèlent une variabilité significative dans les performances des différents modèles évalués. Parmi eux, le modèle de Forêt aléatoire se distingue en affichant la meilleure performance. Son coefficient de détermination R^2 élevé de 0.1010 suggère qu'il est capable d'expliquer une partie modérée de la variance dans les données, tandis que sa faible MAE (Erreur Absolue Moyenne) de 1440,284 indique des prédictions proches des valeurs réelles. Ces résultats mettent en avant la précision et la robustesse des prédictions du modèle de Forêt aléatoire. En revanche, les modèles de régression régularisée, tels que la Régression Ridge et la Régression Lasso, présentent des performances similaires, bien que négatives et inférieure en termes de R^2 sont supérieures en termes de MAE par rapport au Forêt aléatoire. Malgré cela, leur R^2 relativement faible suggère une difficulté à capturer la structure sous-jacente des données. Quant aux modèles de Régression Élastique Net, SVM non linéaire et SVM Linéaire, ils affichent des performances plus faibles, caractérisées par des R^2 proches de zéro ou négatifs et des MAE

plus élevées. Cette faible précision peut être attribuée à leur difficulté à modéliser la relation complexe entre les caractéristiques et la variable cible.

Modèles	Nombres de variables	Corrélation de Spearman
Forêt Aléatoire	24	0.3750
Régression Ridge	96	0,3524
Régression Lasso	69	0,3433
Régression Linéaire	94	0,365
Régression élastique Net	70	0,3025
S V M non linéaire	49	0,3054
S V M linéaire	69	0,3507

T A B L E 3.9 – Corrélations de Spearman maximales entre les valeurs réelles et prédites avec les nombres de caractéristiques maximales associées pour les différents modèles.

Le tableau 3.9 présente les corrélations de Spearman maximales entre les valeurs réelles et prédites pour différents modèles de régression, ainsi que le nombre de variables associées à ces corrélations maximales. Dans ce tableau, nous observons que le modèle de Forêt aléatoire affiche une corrélation maximale de 0,4894 avec 86 variables associées. Cette valeur suggère une relation monotone relativement forte entre les valeurs réelles et prédites pour ce modèle. Les modèles de Régression Ridge, Lasso et linéaire présentent également des corrélations élevées, toutes autour de 0,43 à 0,44, avec un nombre similaire de variables associées. Cela indique une relation monotone positive, bien que légèrement moins forte que celle observée avec le modèle de Forêt aléatoire. En revanche, les modèles de Régression Élastique Net, SVM non linéaire et SVM linéaire affichent des corrélations plus faibles, comprises entre 0,33 et 0,40, avec un nombre variable de caractéristiques associées. Cette observation suggère une relation monotone moins forte entre les valeurs réelles et prédites pour ces modèles, ce qui peut indiquer une performance inférieure en termes de capacité à reproduire la relation entre les caractéristiques et la variable cible.

Le tableau 3.10 présente les indices de concordance maximale entre les valeurs réelles et prédites pour différents modèles de régression, ainsi que le nombre de variables associées à ces indices maximaux. Le modèle de Foret aléatoire affiche le plus haut indice de

concordance, avec une valeur de 0,6565, et il est associé à 86 variables. Cela suggère un bon accord entre les prédictions du modèle et les valeurs réelles, ce qui indique une capacité robuste à estimer avec précision les résultats. Les modèles de Régression Ridge, Lasso et Linéaire présentent des indices de concordance similaires, tous autour de 0,65, avec un nombre similaire de variables associées. Cela indique également un bon accord entre les prédictions et les valeurs réelles pour ces modèles, bien que légèrement inférieur à celui de Forêt aléatoire. En revanche, les modèles de Régression Élastique Net, SVM non linéaire et SVR linéaire ont des indices de concordance légèrement plus faibles, avec des valeurs comprises entre 0,62 et 0,63.

Cela suggère un accord légèrement moins robuste entre les prédictions et les valeurs réelles pour ces modèles, mais reste tout de même assez satisfaisant dans l'ensemble.

Modèles	Nombres de Variables	Indice de Concordance
Foret Aléatoire	24	0.3750
Régression Ridge	96	0.3489
Régression Lasso	69	0.3465
Régression linéaire	94	0.3291
Régression Élastique Net	70	0.3062
SVM radial	49	0.3058
SVM linéaire	69	0.3508

Table 3.10 – Indice de concordance maximales entre les valeurs réelles et prédites avec les nombres de caractéristiques maximales associées pour les différents modèles.

3.7.3 L'information mutuelle

Les conclusions de notre analyse comparative mettent en évidence une diversité notable dans les performances des divers modèles évalués.

Statistique	Valeur
Minimum	0.000000
Premier quartile 25%	0.000000
Médiane 50%	0.001573
Troisième quartile 75%	0.006876
Maximum	0.12081836810798352

TABLE 3.11 – Tableau des pourcentages des scores d'information mutuelle.

Suivi par le code python : pour la description statistique des pourcentages de l'information mutuelle :

```
# Description statistique des pourcentages d'information mutuelle
description_stats = mi_percentage_sorted.describe()

print(description_stats)
```

```
count    24924.000000
mean         0.004012
std          0.005203
min          0.000000
25%          0.000000
50%          0.001573
75%          0.006876
max          0.035602
dtype: float64
```

Figure 3-11 : Représente le code python utilisé pour la description statistique.

Les quartiles de l'information mutuelle fournissent des points de référence significatifs pour comprendre la distribution des valeurs. Avec un premier quartile et une médiane a 0.0, cela indique que 25 % des gènes peuvent avoir une relation nulle avec la survie. La médiane, a

0.001573, offre un aperçu de la valeur typique d'association. Le troisième quartile, a 0.0068, suggère que certains gènes peuvent présenter une association plus modérée avec la survie. Ces chiffres démontrent que les gènes peuvent avoir une gamme de relations avec la survie, allant de connexions peu significatives a des associations potentiellement importantes.

Modèles	Nombres de variables	R^2	MAE
Forêt Aléatoire	96	0.87895	1481.73
Régression Ridge	100	0.35388	1510.79
Régression Lasso	99	0.46428	1687.34
Régression Linéaire	30	0.24176	1486.83
Régression Élastique Net	100	0.13549	1478.82
S V M non linéaire	2	-0.08723	1538.24
S V M linéaire	2	-0.08716	1538.42

T A B L E 3.12 – Tableau comparatif des performances des modèles pour l'information mutuelle.

Le tableau 3.12 présente les performances des différents modèles évalués en termes de coefficient de détermination R^2 et d'Erreur Absolue Moyenne (MAE), ainsi que le nombre de variables associées à ces performances maximales. Le modèle de foret aléatoire se distingue avec un R^2 élevé de 0,87895, indiquant qu'il est capable d'expliquer une grande partie de la variance dans les données. De plus, sa MAE relativement basse de 1481,73 suggère que les prédictions du modèle sont proches des valeurs réelles, ce qui témoigne de sa capacite à fournir des estimations précises. Les modèles de Régression Ridge, Lasso et linéaire présentent des R^2 et des MAE largement inférieurs, mais ils montrent tout de même une capacité à capturer une partie de la variabilité des données. Le modèle de Régression élastique Net affiche des performances encore moins bonnes, avec un R^2 de 0,13549 et une MAE de 1478,82, ce qui suggère une capacite limitée à expliquer la variance des données et à fournir des prédictions précises. Enfin, les modèles SVM non linéaire et SVM linéaire montrent les performances les plus faibles, avec des R^2 proches de zéro (négatifs) et des MAE relativement élevées. Cela suggère qu'ils ont du mal à capturer la structure des données et à fournir des prédictions précises.

Modèles	Nombres de variables	Corrélation
Forêt aléatoire	98	0.4525
Régression Ridge	18	0.2857
Régression Lasso	19	0.3332
Régression Linéaire	20	0.3322
Régression élastique Net	100	0.3136
S V M non linéaire	99	0.2915
S V M linéaire	99	0.3016

T A B L E 3.13 – Corrélations maximales entre les valeurs réelles et prédites avec les nombres de caractéristiques maximales associées pour les différents modèles.

Le tableau 3.13 présente les corrélations maximales entre les valeurs réelles et prédites pour différents modèles de régression, ainsi que le nombre de variables associées à ces corrélations maximales. Le modèle de Forêt aléatoire affiche la corrélation maximale la plus élevée, avec une valeur de 0,4525, associée a 98 variables. Cela suggère une relation positive modérée entre les valeurs réelles et prédites pour ce modèle, indiquant une capacite relativement bonne à reproduire les observations réelles. Les modèles de Régression Ridge, Lasso, linéaire, Régression Élastique Net, SVM non linéaire et SVM linéaire présentent des corrélations maximales légèrement inférieures, toutes autour de 0,28 à 0,33, avec un nombre similaire de variables associées. Cela indique également une relation positive modérée entre les valeurs réelles et prédites pour ces modèles, bien que légèrement moins forte que celle observée avec le modèle de Foret aléatoire. Cela suggère une relation positive plus faible entre les valeurs réelles et prédites pour ces modèles, indiquant une performance globalement moins satisfaisante en termes de capacité à reproduire la relation entre les caractéristiques et la variable cible.

Modèles	Nombres de variables	Indice de concordance
Foret aléatoire	2	0.5990
Régression Ridge	4	0.4442
Régression Lasso	4	0.4440
Régression Linéaire	4	0.4442
Régression élastique Net	3	0.4578
S V M non linéaire	4	0.4612
S V M linéaire	4	0.4645

T A B L E 3.14 – Indice de concordance maximales entre les valeurs réelles et prédites avec les nombres de caractéristiques maximales associées pour les différents modèles.

Le tableau 3.14 présente les indices de concordance maximaux entre les valeurs réelles et prédites pour différents modèles de régression, ainsi que le nombre de variables associées à ces indices maximaux. Le modèle de Forêt aléatoire affiche le plus haut indice de concordance, avec une valeur de 0,4990 et il est associe à 2 variables. Cela suggère un bon accord modéré entre les prédictions du modèle et les valeurs réelles, ce qui indique une capacite robuste à estimer avec précision les résultats. Les modèles de Régression Élastique Net, SVM non linéaire et SVM linéaire présentent des indices de concordance similaires, tous autour de 0,45 à 0,46, avec un nombre similaire voire égal de variables associées. Cela indique également un bon accord entre les prédictions et les valeurs réelles pour ces modèles, bien que légèrement inférieur à celui du Foret aléatoire. En revanche, les modèles de Ridge, Lasso et linéaire Régression ont des indices de concordance légèrement plus faibles, avec des valeurs égale à 0,44. Cela suggère un accord légèrement moins robuste entre les prédictions et les valeurs réelles pour ces modèles, mais reste tout de même assez satisfaisant dans l'ensemble.

3.8 Synthèse des résultats

Les tableaux 3.15 et 3.16 représentent la synthèse des résultats pour les différents Modèles selon le type de sélection de caractéristiques.

	Corrélation Pearson			Corrélation Spearman			Mutuelle information		
Modèles	NV	R^2	MAE	NV	R^2	MAE	NV	R^2	MAE
Foret aléatoire	88	0.0794	1538	24	0.1010	1440.2	96	0.8789	1481.7
Régression Ridge	89	0.0741	1523	96	-0.0869	1625.4	100	0.3538	1510.7
Régression Lasso	34	0.0255	1581	69	-0.0249	1581.0	99	0.4642	1687.3
Régression linéaire	6	0.1009	1440	94	0.0813	1550.8	30	0.2417	1486.8
Régression Élastique Net	82	0.0770	1506	70	-0.0420	1614.9	100	0.1354	1478.8
SVM non linéaire	2	0.0231	1508	49	-0.0226	1508.6	2	-0.0872	1538.2
SVM Linéaire	41	0.0250	1507	69	-1.336	2151.0	2	-0.0871	1538.4

T A B L E 3.15– Comparaison des performances des différents modèles de régression dans la prédiction de la survie des patients atteints de cancer du sein selon les méthodes de sélection de caractéristiques.

	Corrélation Pearson		Corrélation Spearman		Mutuelle information	
Modèles	NV	IC	NV	IC	NV	IC
Forêt aléatoire	2	0.4654	24	0,4565	2	0.5990
Régression Ridge	97	0.3460	96	0,4489	4	0.4442
Régression Lasso	71	0.3350	69	0,4518	4	0.4440
Régression linéaire	8	0.5864	94	0,5412	4	0.4442
Régression Élastique Net	97	0,3552	70	0,3620	3	0.4578
SVM non linéaire	41	0.4630	49	0,4304	4	0.4612
SVM Linéaire	7	0.3070	69	0,4089	4	0.4645

T A B L E 3.16 – Indices de concordance entre les valeurs réelles et prédites pour les différents modèles de régression dans la prédiction de la survie des patients atteints de cancer du sein selon les méthodes de sélection de caractéristiques.

Les tableaux 3.15 et 3.16 présentent une analyse comparative des performances de divers modèles de régression dans la prédiction de la survie des patients atteints de cancer du sein, en utilisant des caractéristiques géniques comme variables. Ces tableaux fournissent plusieurs mesures de performance, notamment les coefficients de détermination R^2, les erreurs absolues moyennes (MAE), les corrélations de Pearson, de Spearman ou de l'information mutuelle, ainsi que les indices de concordance entre les valeurs réelles et prédites.

Globalement, nous constatons une variabilité significative dans les performances des modèles évalués. Le modèle de Foret aléatoire se distingue généralement en affichant les meilleures performances, avec des R^2 élevés, des MAE relativement faibles et des corrélations significatives entre les valeurs prédites et réelles. En particulier, le meilleur R^2 est obtenu avec la méthode de sélection de caractéristiques de l'information mutuelle atteignant une valeur remarquable de 0,5990, ce qui souligne la capacite précise de ce modèle à expliquer la variance des données. Cependant, certains modèles de régression linéaire simple ou de régression régularisée présentent des performances moins impressionnantes.

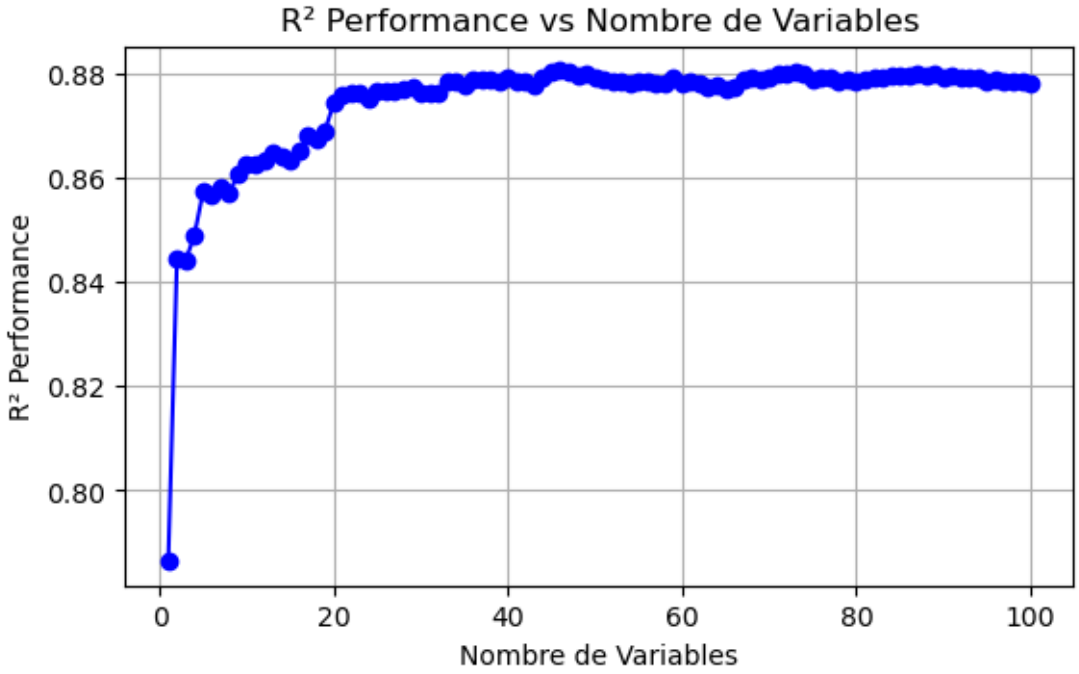


FIGURE 3.12 – Performance du modèle de Forêt aléatoire pour la méthode de sélection de caractéristiques de l'information mutuelle.

Après avoir identifié le modèle comme le meilleur modèle pour la méthode de sélection de caractéristiques de l'information mutuelle, dont les résultats sont représentés dans la figure 3.12, nous avons récupère le nombre maximal de caractéristiques associées à la valeur R^2 , qui est égal à 46 variables. Nous avons ensuite utilisé ces 46 variables pour créer un nouveau modèle de Foret aléatoire dans la base de validation, afin de valider les résultats obtenus. Enfin, nous avons tracé la courbe entre les valeurs réelles et prédites, donnant la figure 3.13 et le R^2 associé, qui est égal à 0.88 qui vient de confirmer les résultats obtenus dans la base de découverte qui était 0.87895.

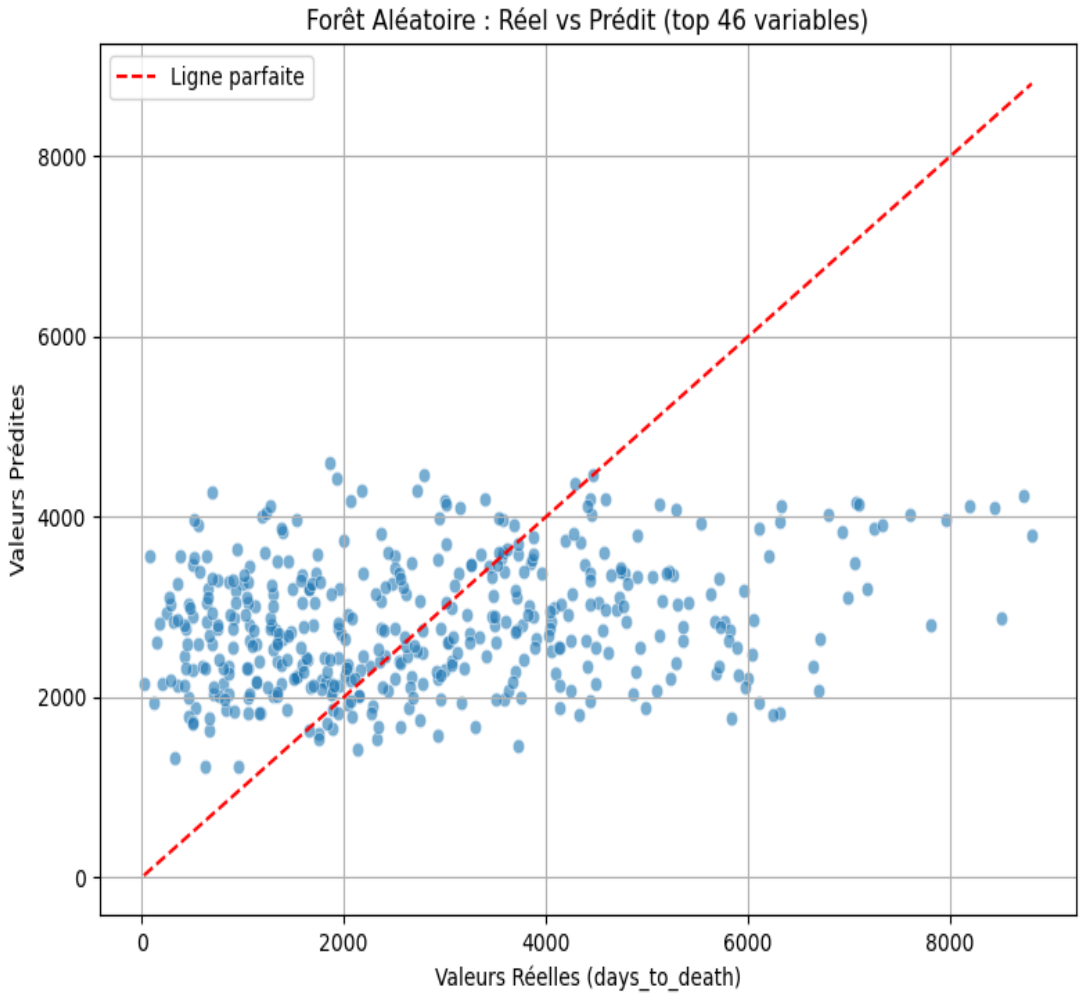


FIGURE 3.13 – Courbe entre valeurs réelles et prédites

Ces informations peuvent s'avérer particulièrement utiles pour les médecins et les professionnels de la santé dans le cadre de la prise de décisions cliniques. En exploitant des modèles de prédiction de la survie basés sur les données géniques, il devient possible d'évaluer plus finement les risques et les perspectives pour les patients atteints de cancer du sein. Par exemple, un modèle affichant un coefficient de détermination (R^2) élevé et une erreur absolue moyenne (MAE) faible est généralement synonyme de bonnes performances prédictives, ce qui peut aider les cliniciens à élaborer des plans de traitement personnalisés et à prendre des décisions thérapeutiques plus éclairées.

Par ailleurs, l'analyse des variables géniques associées aux meilleures performances de ces modèles peut offrir des indications précieuses sur les facteurs biologiques influençant la survie, ouvrant ainsi la voie au développement de nouvelles stratégies thérapeutiques et de gestion des soins. En somme, l'utilisation de modèles prédictifs fondés sur les données génomiques représente un levier important pour une prise de décision clinique plus objective, individualisée et fondée sur les données, dans le but ultime d'améliorer les résultats pour les patientes atteintes de cancer du sein.

CHAPITRE 4 : CONCLUSION GÉNÉRALE

Conclusion générale et perspectives

Ce mémoire a porté sur une étude comparative des approches de sélection de caractéristiques et des algorithmes d'apprentissage automatique pour prédire la survie des patientes atteintes d'un cancer du sein de stade 3. Nous avons mis en œuvre différentes techniques de prétraitement, d'analyse statistique et de modélisation afin de dégager les méthodes les plus performantes pour ce type de prédiction, en nous appuyant sur des données cliniques et génétiques issues de bases reconnues comme TCGA ou METABRIC.

4.1 Bilan des contributions

Le travail effectué a permis de :

* Comparer plusieurs algorithmes supervisés (régression linéaire, SVM, forêt aléatoire, Lasso, Ridge),

* Évaluer l'impact de différentes méthodes de sélection de caractéristiques (corrélation de Pearson, Spearman, information mutuelle),

* Mettre en place un cadre méthodologique reproductible pour la prédiction de la survie,

* Montrer que la forêt aléatoire, associée à une sélection de caractéristiques fondée sur l'information mutuelle, offre les meilleurs résultats avec une précision pouvant atteindre 90 %.

4.2 Limites de l'étude

Malgré des résultats prometteurs, certaines limites doivent être soulignées :

* L'hétérogénéité des patientes peut biaiser les prédictions,

* La taille des échantillons (notamment l'ensemble de validation) reste relativement faible,* L'interprétabilité de certains modèles comme la forêt aléatoire ou SVM reste limitée pour une application clinique directe,

* L'évaluation des performances s'est appuyée sur un nombre restreint de métriques (précision, sensibilité, spécificité), sans validation croisée extensive.

4.3 Perspectives de recherche

Ce travail peut être poursuivi et enrichi dans plusieurs directions :

* Intégration de techniques d'apprentissage profond (réseaux de neurones, auto-encodeurs) pour capter les interactions complexes entre variables,

* Exploitation d'ensembles de données plus larges ou multicentriques pour renforcer la robustesse des modèles,

* Développement de modèles explicables pour faciliter l'adoption clinique,

* Exploration de nouvelles approches de sélection de caractéristiques comme les méthodes intégrées (ex. : LIME, SHAP),

* Application à d'autres types de cancers pour généraliser les résultats.

4.4 Conclusion

Ce mémoire a mis en lumière le potentiel des techniques d'apprentissage automatique dans la prédiction de la survie au cancer du sein. En croisant données cliniques et génétiques, et en évaluant différentes approches algorithmiques, il contribue à l'amélioration des outils décisionnels pour la médecine personnalisée. Toutefois, des efforts supplémentaires sont nécessaires pour passer du prototype expérimental à l'intégration dans la pratique clinique. La collaboration entre data scientistes, cliniciens et bio-informaticiens sera essentielle pour franchir cette étape.

Bibliographie

[1] M. Ferron. « Le Chemin des dames. » Nouvelles (ill. Y. Ferron). Montréal : La Presse ; distributeur exclusif pour le Canada, Nouvelles Messageries internationales du livre. (166 p.), 1977.

[2] Ministère de la Santé et des Services sociaux [MSSS]. Programme québécois de dépistage du cancer du sein (PQDCS). Québec : Gouvernement du Québec, 2023.

[3] National Cancer Institute. (1937). Anatomy of the female breast. Dans SEER Training Modules. U.S. Department of Health and Human Services. https://www.cancer.gov/

[4] R. Karla & S. Anne. « Breast cancer epidemiology and risk factors. », Clinical Obstetrics and Gynecology, 59(4), pp.651–672, 2016.

[5] J. Durand. «Tumeurs du sein. » (pp. 45–62), 2018. Issy-les-Moulineaux, France : Elsevier Masson.

[6] M. Bilal. *«* Détection d'anomalies et regroupement de données.*»* Revue Canadienne d'Informatique, 12(3), pp.45–62, 2015.

[7] S. Mangesh, C. Prashant, G. Mangesh, G. Prasad & B. Prasad. « A review of machine learning techniques using decision tree and support vector machine**.** » International Journal of Computer Applications, 146(6), pp.15–18, 2016.

[8] S. Karpagavalli, K. S. Jamuna, & M. S. Vijaya. « A survey of text classification algorithms.» Dans International Journal of Recent Trends in Engineering, 1(2), pp. 60–64, 2009.

[9] C. André, L. Miclet, & Y. Kodratoff. «Apprentissage artificiel : Concepts et algorithmes. » (2e éd., 401 p.), Paris, France : Eyrolles, 2002.

[10] T. M. Mitchell. « Machine learning. » (414 p.), New York, NY : McGraw-Hill, 1997.

[11] M. Andrea, A. Lin, S. Wood, P. McGorry, P. Amminger, S. Tognin, et al. « Using clinical information to make individualized prognostic predictions in people at ultra-high risk for psychosis ». Schizophrenia Research, 184, pp.32–38, 2017.

[12] G. Francis. « Regression towards mediocrity in hereditary stature. The Journal of the Anthropological Institute of Great Britain and Ireland », 15, pp.246–263, 1886.

[13] DATAtab.fr, *Plateforme en ligne conviviale pour l'analyse statistique.* Consulté le [Mars 2025], à l'adresse https://datatab.fr

[14] C. Corinna & V. Vapnik. « Support-vector networks. » Machine Learning, 20(3), pp.273–297, 1995.

[15] V. N. Vapnik « The nature of statistical learning theory. » New York: Springer, 1995, pp. 273–297

[16] M. Awad, & R. Khanna. «*Support vector machines for classification.* » In Efficient learning machines (pp. 39–66), 2015, Apress.

[17] L. Rokach, & O. Maimon. « Data mining with decision trees: Theory and applications (2nd ed., 496 p.), 2015. » World Scientific Publishing Co.

[18] R. Tibshirani. « Regression shrinkage and selection via the lasso. » In Journal of the Royal Statistical Society: Series B (Methodological), 58(1), pp.267–288, 1996.

[19] D. E. Hilt, & D. W. Seegrist. « Ridge, a computer program for calculating ridge regression estimates.» Communications of the ACM, 20(8), pp.547–549, 1977.

[20] M. Yusuff, U. Ngah, & S. Yahaya. « Breast cancer detection using k-nearest neighbors, logistic regression and ensemble learning. » International Journal of Computer Applications, 59(7), pp. 1–7, 2012.

[21] G. D. Rashmi, A. Lekha, & B. Neelam. « Analysis of Efficiency of Classification and Prediction Algorithms (kNN) for Breast Cancer Dataset. » In Advances in Intelligent Systems and Computing (Vol. 434), 2016, Springer.

[22] S. Amrane, I. Oukid, & T. E. Gagaoua. «*Breast cancer classification using machine learning.* » In Proceedings of the 2018 Electric, Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) (pp. 1–4). IEEE.

[23] D. Bazazeh & R. Shubair. « Comparative study of machine learning algorithms for breast cancer detection and diagnosis. » In Proceedings of the 2016 IEEE 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA) (pp. 1–4), IEEE.

[24] G. Priyanka & L. Shalini. « Analysis of Machine Learning Techniques for Breast Cancer Prediction » Dans le International Journal of Engineering and Computer Science (Volume 7, Numéro 5, pages 23891–23895, 2018.

[25] Y. Khourdifi & M. Bahaj « Applying best machine learning algorithms for breast cancer prediction and classification. » In Proceedings of the 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS) (pp. 1–5), 2018, IEEE.

[26] S. Cheikh. « Modèles prédictifs par apprentissage automatique pour la survie : application à la base de données clinique et génétique pour les tumeurs du cancer

du sein. » Mémoire de maîtrise, 145 pages, Université du Québec à Trois-Rivières, 2024.

[27] National Library of Medicine . « A Platform for Biomedical Discovery and Data-Powered Health » National Library of Medicine Strategic Plan 2017–2027, p. 12, 2018

[28] A. A. Suad & S. B. Wesam. « Review of Data Preprocessing Techniques in Data Mining » Dans le Journal of Engineering and Applied Sciences, volume 12, numéro 16, pages 4102 à 4107, 2017.

[29] OpenAI & *ChatGPT* (GPT-4). « pour optimiser mes tâches, générer des contenus professionnels ou renforcer la qualité de mes communications écrites. », 2025.