REPUPLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Ministère De L'enseignement Supérieure De La Recherche Scientifique

Université de KASDI MERBAH - Ouargla

Faculté des mathématiques et des sciences de la matière

Département de physique : Domaine science de la matière

Mémoire

Pour l'obtention du diplôme de

Master professionnel en physique médicale

Présenté par :

- **BERGUIGA Intisar**
- **LAIB Soumia**

Thème :

Apprentissage automatique pour la prédiction des quelques traitement du cancer du sein dans la région d'Ouargla.

Soutenu publiquement 10/06/2024 devant le jury composé de :

BENHAMIDA Soufiane	MCA	Université de Kasdi Merbah Ouargla	Président
REMITA Naamane	MCB	Université de Kasdi Merbah Ouargla	Examinateur
AYAT Zahia	MCA	Université de Kasdi Merbah Ouargla	Encadrent
AIADI Oussama	MCA	Université de Kasdi Merbah Ouargla	Co-Encadreur

Année Universitaire : 2023/2024

Remercîment

On remercie dieu le tout puissant de nous avoir donné la santé et la volonté d'entamer et de terminer ce mémoire.

Tout d'abord, ce travail ne serait pas aussi riche et n'aurait pas pu avoir le jour sans l'aide et l'encadrement de ***Dr. AYAT Zahia,*** *pour sa patience, et sa disponibilité durant notre préparation de ce mémoire, et au Co-Encadreur* ***Dr. AIADI Oussama.***

Nous tenons à remercier le service ***CAC OUARGLA et Dr. DEBBA Issam*** *et* ***Mr. Khadraoui*** *pour les installations nécessaires à la réalisation de nos travaux.*

Nos remerciements sont également adressés aux membres de jury qui ont accepté de juger ce travail : ***Dr. BENHAMIDA Soufiane*** *et* ***Dr. REMITA Naamane***

Enfin, Nous exprimons également notre gratitude à tous les professeurs et assistants de l'Université de Ouargla en général, et particulièrement du Département de Physique de la Faculté de Mathématiques et Sciences des Matériaux, pour leur dévouement.

Dédicace

Merci mon dieu de m'avoir donné la capacité de patience d'aller jusqu'au bout du rêve J'ai l'honneur de dédier ce modeste travail :

A ma mère Assia qui s'est sacrifiée pour mon bonheur et ma réussite.

A mon père Azzouzi qui a été mon ombre durant toutes les années des études.

A mes frères, mon soutien dans la vie: Alla, Chawki, Anis , Samah, Aya et ma belle soeur Nesrine

Je n'oublierai pas les compagnons d'âme qui ont partagé avec moi les étapes de ce chemin, ceux qui m'ont encouragé à poursuivre la marche, je vous en suis reconnaissant à tous.

Intisar

Dédicace

Louange à Dieu seul Le tout puissant de nous avoir donné la santé et la volonté d'entamer et de terminer ce mémoire.

Le but était grand, le chemin était facile et le plaisir que l'arrivée venait pour effacer la fatigue des années.

Ce modeste travail est dédié spécialement :

À celui qui m'a appris à donner sans attendre, à celui dont je porte son nom avec fierté. «Mon cher père que Dieu t'accepte avec ses martyrs ».

À ma chère mère, ma raison de vivre en témoignage de la reconnaissance pour sa patience, son amour et ses sacrifices.

Pour mon réconfort dans la vie et tout au long de mon chemin mes chers frères Hocine et Mouad et ma princesse (ma sœur) Amina.

À ceux qui portaient des souvenirs dans leur cœur, et à ceux qui ont partagé avec moi mes moments doux et amers, mes amis : Dounia, Achouak , Hafsa, Aya.

Au compagnon des premiers pas, à celui qui se tenait derrière moi comme une ombre, à celui qui m'a aidé dans mes moments difficiles, ma compagne de mémoire : Intissar

Au baume de l'âme et la vie, le symbole de la fidélité : Salsabil .

Aux mes camarades de la promotion de physique médicale 2022/2024.

À tous ces qui ont contribué à ma réussite par un mot ou une prière : ma famille et ma belle-sœur Naima Louange à Allah.

Soumia

Résume :

Le cancer du sein est l'un des types de cancer les plus courants chez la femme, et dans ce travail, nous souhaitons proposer une nouvelle approche basée sur une branche de l'intelligence artificielle, à savoir l'apprentissage supervisé. Il vise à aider les médecins à diagnostiquer et à traiter les patients. Les données des patients étudiés ont été collectées à partir de l'archive de Centre Anti Cancéreux (CAC) de l'EPH Mohamed Boudiaf Ouargla.

Pour contribuer à l'hormonothérapie en utilisant une méthode de classification, nous avons comparé les performances de différents algorithmes en fonction de la précision et de la sensibilité du modèle. Les résultats expérimentaux ont montré que le pourcentage le plus élevé a été obtenu en appliquant une régression logistique et Bayes naïf gaussien avec une précision de 91 %.

D'autre part, il a été prévu que la dose totale trouverait le meilleur modèle en appliquant une régression et en comparant les métriques d'évaluation des algorithmes, où la précision du modèle à noyau linéaire était bonne, équivalente à 98,5 %.

Mots clés : intelligence artificielle, apprentissage automatique, cancer du sein, dose.

ملخص:

يعد سرطان الثدي أحد أكثر أنواع السرطان شيوعًا بين النساء، وفي هذا العمل، نود أن نقترح نهجًا جديدًا يعتمد على فرع من فروع الذكاء الاصطناعي، وهو التعلم الخاضع للإشراف. يهدف الى مساعدة الاطباء في تشخيص وعلاج المرضى. تم جمع بيانات المرضى التي أجريت عليها الدراسة من أرشيف قسم العلاج الإشعاعي محمد بوضياف بورقلة.

للمساهمة في العلاج الهرموني باستخدام طريقة التصنيف، لتحقيق ذلك قمنا بمقارنة أداء الخوارزميات المختلفة على أساس دقة وحساسية النموذج. وأظهرت النتائج التجريبية أن أعلى نسبة تم الحصول عليها بالتطبيق الانحدار اللوجستي و غاوسي بايز. بدقة 91.%

من ناحية اخرى تم التنبؤ بالجرعة الإجمالية لإيجاد افضل نموذج من خلال تطبيق الانحدار ومقارنة مقاييس التقييم للخوارزميات حيث كانت دقة نموذج النواة الخطية جيدة ،تعادل 98.5%

الكلمات المفتاحية : الذكاء الاصطناعي ، التعلم الالي ، سرطان الثدي ، الجرعة .

Abstract:

Breast cancer is one of the most common types of cancer in women, and in this work, we would like to propose a new approach based on a branch of artificial intelligence, namely supervised learning. It aims to help doctors diagnose and treat patients. The data of the patients studied were collected from the Anti-Cancer Center (CAC) archive of the EPH Mohamed Boudiaf Ouargla.

To contribute to hormonal therapy using a classification method, we compared the performance of different algorithms based on the accuracy and sensitivity of the model. The experimental results showed that the highest percentage was obtained by applying linear regression and Gaussian NB with an accuracy of 91%.

On the other hand, the total dose was predicted to find the best model by applying regression and comparing evaluation metrics for the algorithms, where the accuracy of the linear kernel model was good, equivalent to 98.5%.

Keywords: artificial intelligence, machine learning, breast cancer, dose.

CONTENU :

CHAPITER II

CHAPITER III

LIST DES FIGURE

Chapitre I

Chapitre II

Chapitre III

LIST DES TABLEAUX :

Chapitre I

Chapitre II

Chapitre III

LISTE DES ABREVIATIONS :

ADN: Acide Désoxyribonucléique.

CAC: Centre Anti Cancer.

CCI: Carcinome Canalaire Infiltrant.

CCIS: Carcinome Canalaire In Situ.

CLI: Carcinome Lobulaire Infiltrant.

CLIS: Carcinome Lobulaire In Situ.

CS: Cancer du Sein.

DL: Deep Learning.

DT: Desision tree.

EPH: L'Etablissement Public Hospitalier.

FN: Faux Négatifs.

FP: Faux Positifs.

GNB : Gaussian Naïve Bayes.

Gy: Gray.

HCA: Hierarchical Clustering.

HER2: Human Epidermal Growth Factor Receptor-2.

IA: L'Intelligence Artificielle.

IRM: L'imagerie par résonance magnétique.

KNN: K-Nearest Neighbors.

MAE: Mean Absolut Error.

ML: Machine Learning.

PCA: Principal Component Analysis.

PPS: Programme Personnalisé de Soins.

R^2: R Squared Error.

RBF: Radial Basis Function.

RCP : Réunion de Concertation Pluridisciplinaire.

REPTree: Decision Tree Classifier.

RMSE: Root Mean Squared Error.

SD/ **SG:** Sein Gauche /Droit.

SSE: Erreur sur la Somme des Carrés.

SST: Somme des Carrés Total.

SVM: Support Vector Machine.

SVR: Support Vector Regression.

TNM: Tumeur, Node, Métastase.

VN: Vrais Négatifs.

VP: Vrais Positifs.

Introduction Générale

Le fardeau du cancer continue de s'alourdir à l'échelle mondiale, exerçant une énorme pression physique, émotionnelle et financière sur les personnes, les familles, les communautés et les systèmes de santé. Dans les pays à revenu faible ou intermédiaire, bon nombre de systèmes de santé sont moins bien préparés à gérer ce fardeau, et partout dans le monde, beaucoup de patients atteints de cancer n'ont pas accès à un diagnostic et à un traitement de qualité en temps utile. Dans les pays où les systèmes de santé sont solides, de nombreux cancers obtiennent de meilleurs taux de survie grâce à un dépistage précoce accessible, à un traitement de qualité et aux soins proposés aux patients ayant réchappé à la maladie.

Le cancer du sein est le type de cancer le plus courant chez les femmes, et il est la deuxième cause de décès par le cancer chez les femmes après le cancer des poumons [1]. Cependant, la détection précoce du cancer du sein permet une réduction significative de la mortalité, ainsi que la sensibilisation aux symptômes et aux signes.

Ces dernières années, le domaine de L'intelligence artificielle et les techniques d'apprentissage automatique ont joué un rôle majeur dans des domaines médicaux. Pour améliorer la précision et l'efficacité de la détection du cancer du sein et déterminer le traitement approprié, les chercheurs se sont tournés vers des technologies avancées telles que l'apprentissage profond et l'apprentissage automatique.

Ce mémoire est organisé comme suit :

- Chapitre I: Le Cancer du sien : Dans ce chapitre, nous allons présenter également les statistiques relatives à cette maladie en Algérie et Ouargla .Ensuite, nous allons défini le cancer du sein, ses symptômes et ses méthodes de traitement.
- Chapitre II: Apprentissage automatique et Apprentissage profond : Au début de ce chapitre, nous allons fournir un regard sur les techniques l'intelligence artificielle ses branches et ses algorithmes.
- Chapitre III: Expérimentations et étude des performances : Ce chapitre se compose de: D'abord, étude statistique des donner des patients. Puis, Nous allons, appliques les algorithmes de apprentissage supervisé (classification, regression), et fait une comparaison des résultats de test et résultats de entrain.

Chapitre I :

Généralité sur le cancer du sein

I.1.Introduction :

Les cancers naissent de cellules au départ saines et fonctionnelles qui sont devenues anormales suite à l'accumulation d'altérations dans leur patrimoine génétique (ADN) [2].

Au fur et à mesure du temps, les cellules cancéreuses continuent à accumuler. Elles acquièrent ainsi de nouvelles propriétés qui leur permettent de se développer. Détours les tissus de l'organe dans lequel elles sont nées, puis par atteindre les tissus voisins .Par ailleurs, certaines cellules tumorales peuvent devenir mobiles, se détacher de la tumeur et migrer à travers les systèmes sanguin ou lymphatique pour former une tumeur secondaire ailleurs dans l'organisme.

Les décès par cancer sont surtout dus aux dommages causés par les métastases. C'est pourquoi il est important de diagnostiquer précocement la maladie, avant sa dissémination dans l'organisme. Le cancer du sein est le type de cancer le plus fréquemment diagnostiqué chez les femmes du monde. Dans ce chapitre, nous fournirons un aperçu complet du cancer du sein et ses méthodes de diagnostic.

I.2.Définition du cancer :

Le cancer est une maladie provoquée par la transformation de cellules qui deviennent anormales et prolifèrent de façon excessive ceci est illustré dans la figure I.1.

Ces cellules déréglées finissent parfois par former une masse qu'on appelle tumeur maligne. Les cellules cancéreuses ont tendance à envahir les tissus voisins et à se détacher de la tumeur initiale. Elles migrent alors par les vaisseaux sanguins et les vaisseaux lymphatiques pour aller former une autre tumeur (métastase) [1].

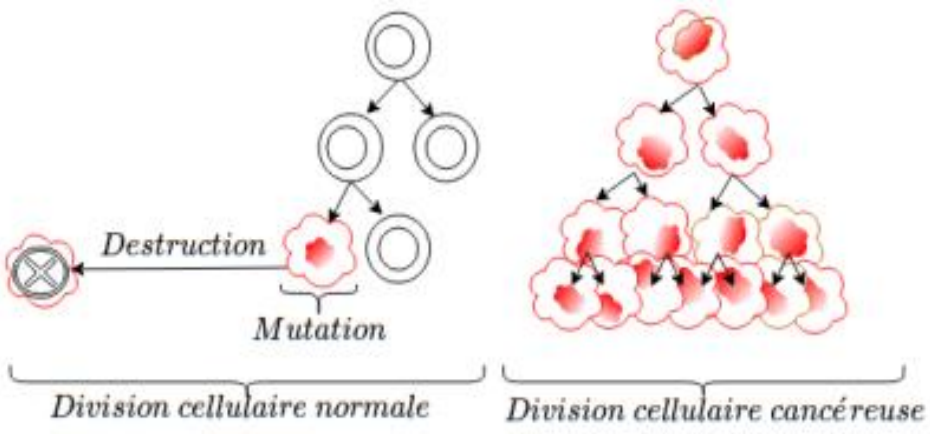


Figure I.1: Cancer division cellulaire [1].

Les cancers du poumon, de la prostate, colorectal, de l'estomac et du foie sont les types de cancer les plus courants chez les hommes, tandis que les femmes sont le plus souvent atteintes des cancers du sein, colorectal, du poumon, du col de l'utérus et de la thyroïde.

I.3.Incidence du cancer du sein :

Le cancer du sein chez les femmes représente l'un des grands problèmes de santé à travers le monde. La mortalité par cancer du sein a peu évolué entre les années 1930 et les années 1970, période pendant laquelle la chirurgie était le mode primaire exclusif de traitement (mastectomie radicale). Le taux de survie a commencé à s'améliorer pendant les années 1990 lorsque des pays ont mis en œuvre des programmes de détection précoce associés à des programmes de traitement complets, avec des thérapies médicales efficaces [1]. À la fin de 2022, 7,8 millions de femmes en vie s'étaient vues diagnostiquer un cancer du sein au cours des cinq années précédentes, ce qui en fait le type de cancer le plus courant à l'échelle du globe [1].

I.3.1.En Algérie

Les dernières statistiques en Algérie montrent quelques résultats concernant le cancer du sein (Figure I.2, I.3). Selon les données du Registre National du Cancer en 2022, le pourcentage total des cas en Algérie est de 22,6% pour les deux sexes, alors que le pourcentage de cancer du sein est de 41,3% [3].

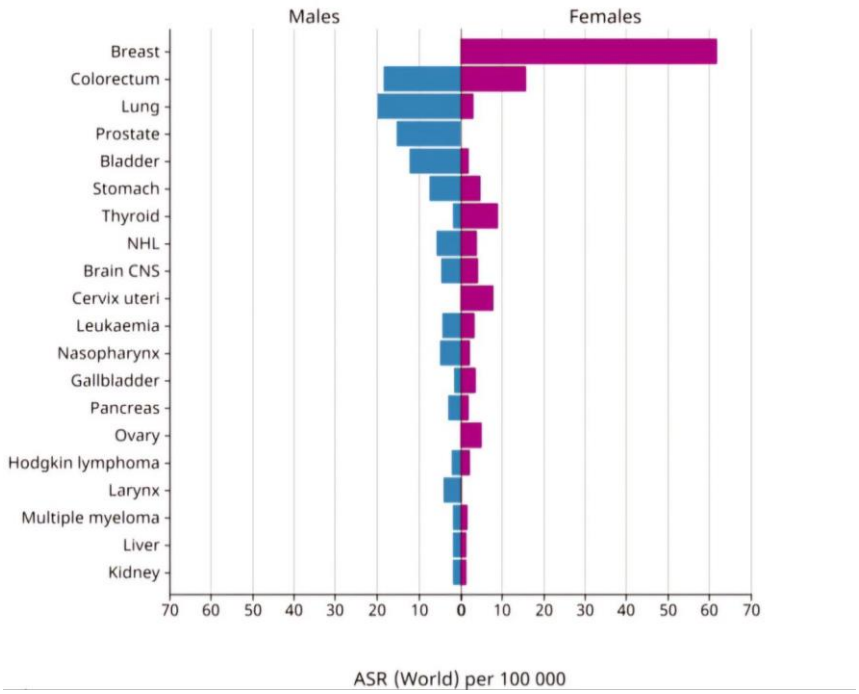


Figure I.2: Chiffres absolus incidence les deux sexes en 2022 Algérie[3].

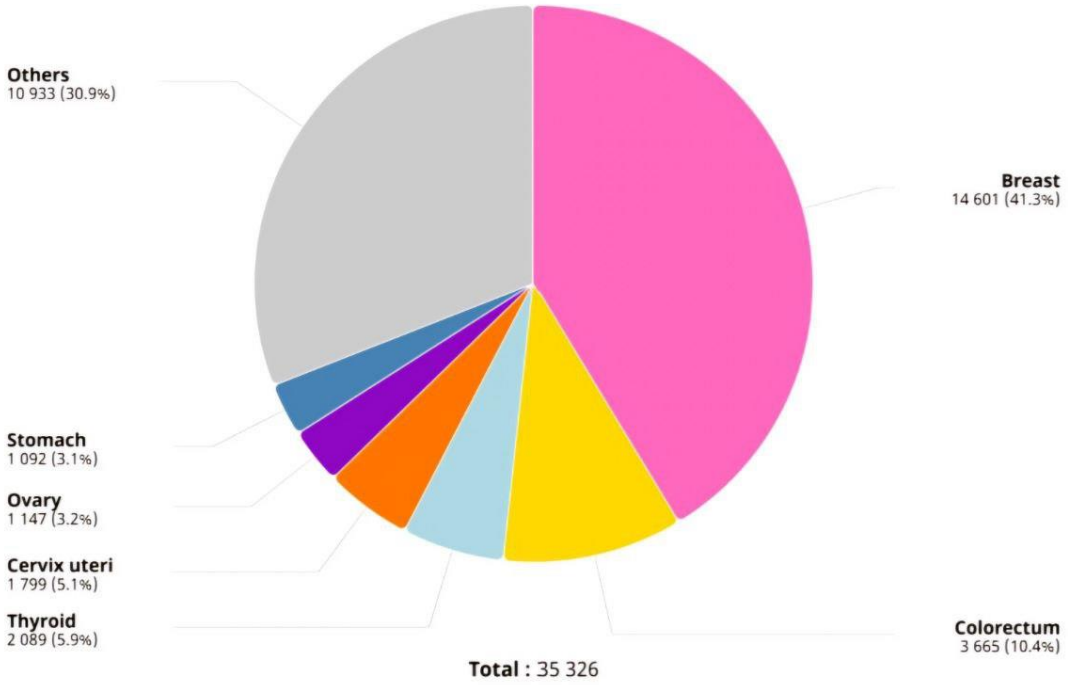


Figure I.3:Taux standardisé selon l'âge (monde) pour 100000 habitants, incident, hommes et femme, en 2022 Algérie [3].

I.3.2.En Ouargla:

Le cancer du sein représente 45% des cancers féminins diagnostiqués durant la période 2015-2019 avec 41,3% des décès par cancer chez la femme [3].

I.4.Définition de cancer du sein :

Le cancer du sein est le cancer le plus fréquent chez la femme. Il se développe à partir des cellules initialement normales qui constituent la glande mammaire. Ces cellules se transforment et se multiplient de façon anarchique et excessive par des mutations ou des instabilités génétiques (anomalie cytogénétique) pour former une masse cellulaire appelée tumeur maligne [4].

I.4.1.Anatomie du sein :

Le sein est situé au niveau de la cage thoracique, sur le muscle grand pectoral qui s'étend de la clavicule au sternum et est lié au creux axillaire. Comme le montre la figure I.4, le sein constitués d'un tissu adipeux qui leur donne leur forme et leur volume, et il évolue selon le naturel des femmes, leur âge, leurs grossesses antérieures, leurs activités physiques… [5] Le sein est constitué de :

► **Les ligaments** qui sont des bandes serrées de tissu conjonctif soutenant les seins. Ils traversent le sein de la peau jusqu'aux muscles où ils se fixent au thorax [6].

► **Les lobules** qui sont des groupes de glandes qui produisent le lait. Chaque sein comporte de 15 à 25 lobules. Les glandes produisent du lait quand elles sont stimulées par les hormones de la femme durant la grossesse [7].

► **Les canaux lactifères** qui sont des tubes qui transportent le lait des lobules au mamelon.

► **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. De taille variable selon les individus, elle a en général un diamètre de 3 cm, mais peut recouvrir entièrement la surface du sein, ou être à peine visible. L'aréole est recouverte de quelques glandes aréolaires qui protègent le sein contre les infections et le dessèchement. L'aréole contient également un muscle qui permet l'érection du mamelon [8].

► **Le mamelon** (téton) désigne la région située au centre de l'aréole et d'où sort le lait à une extrémité. Il est aussi appelé "ostium papillaire" ou "papille". Le mamelon est fait de fibres musculaires. Quand ces fibres se contractent, le mamelon durcit, ou pointe vers l'extérieur. La peau du mamelon est particulièrement fine afin de laisser passer le lait maternel au moment de l'allaitement [8].

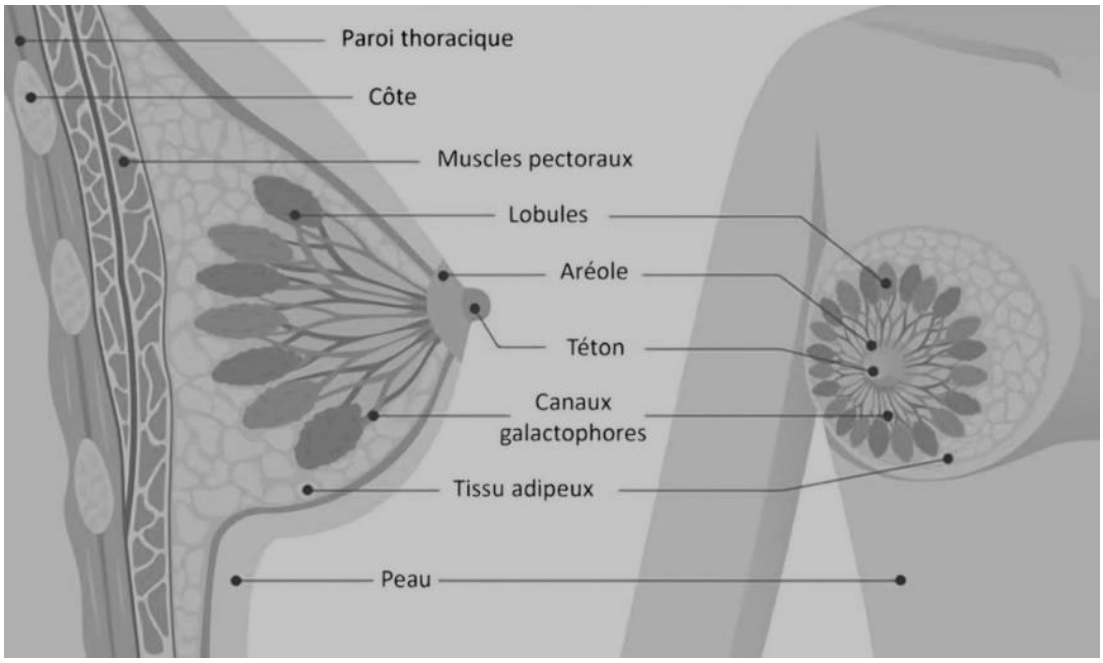


Figure I.4: Anatomie du sein.

I.5.Symptômes du cancer de sein :

Il est possible que le cancer du sein ne cause aucun signe ni symptôme aux tout premiers stades de la maladie [9].

Les symptômes apparaissent quand la tumeur au sein est suffisamment grosse pour qu'on sente la masse au toucher ou quand le cancer s'est propagé aux tissus et aux organes voisins. D'autres affections médicales peuvent causer les mêmes symptômes que le cancer du sein [10].

Le symptôme le plus fréquent du carcinome canalaire est une masse ferme ou dure qui est très différente du reste du tissu mammaire. Elle peut sembler fixée à la peau ou au tissu mammaire voisin. La masse ne rétrécit pas ou ne disparaît pas et ne réapparaît pas au cours du cycle menstruel. Elle peut être sensible mais n'est généralement pas douloureuse (la douleur est plus souvent le symptôme d'une affection non cancéreuse) [11], et les autres symptômes du cancer du sein canalaire ou lobulaire sont illustré dans la Figure I.5.

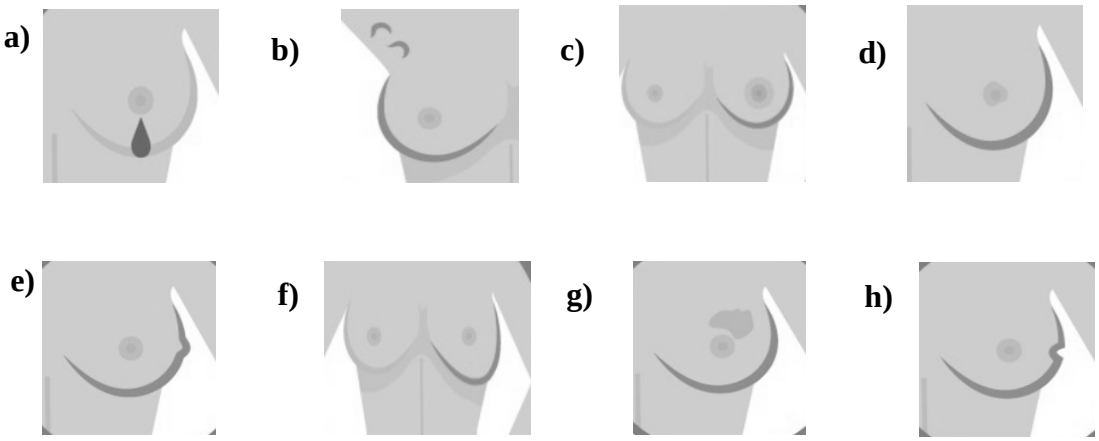


Figure I.5: Les Symptômes du cancer du sein. (a-Écoulement du mamelon sans qu'on le comprime ou qui est teinté de sang. b-Changements mamelonnaires, comme un mamelon qui commence soudainement à pointer vers l'intérieur (mamelon inversé) [12], c-Masse à l'aisselle (creux axillaire), d-Changement d'aspect du mamelon, e-Boule/masse dans le sein, F-Changement de la taille ou de la forme du sein, g-Rougeur, h- Fossette).

Lorsque le cancer du sein s'est propagé au reste du corps, d'autres symptômes peuvent apparaître comme fatigue, nausées, perte de poids, douleurs des os ou troubles de la vision [11,13].

I.6.Les types et stades de cancer du sein :

Il existe différents types de cancer du sein qui est définis au niveau histologique. On peut sont diviser en deux grandes catégories : le cancer non invasif ou in situ, qui touche uniquement les canaux galactophores, et le cancer invasif ou infiltrant, qui se propage dans le tissu gras du sein, ceci est montré sur la Figure I.6 [14].

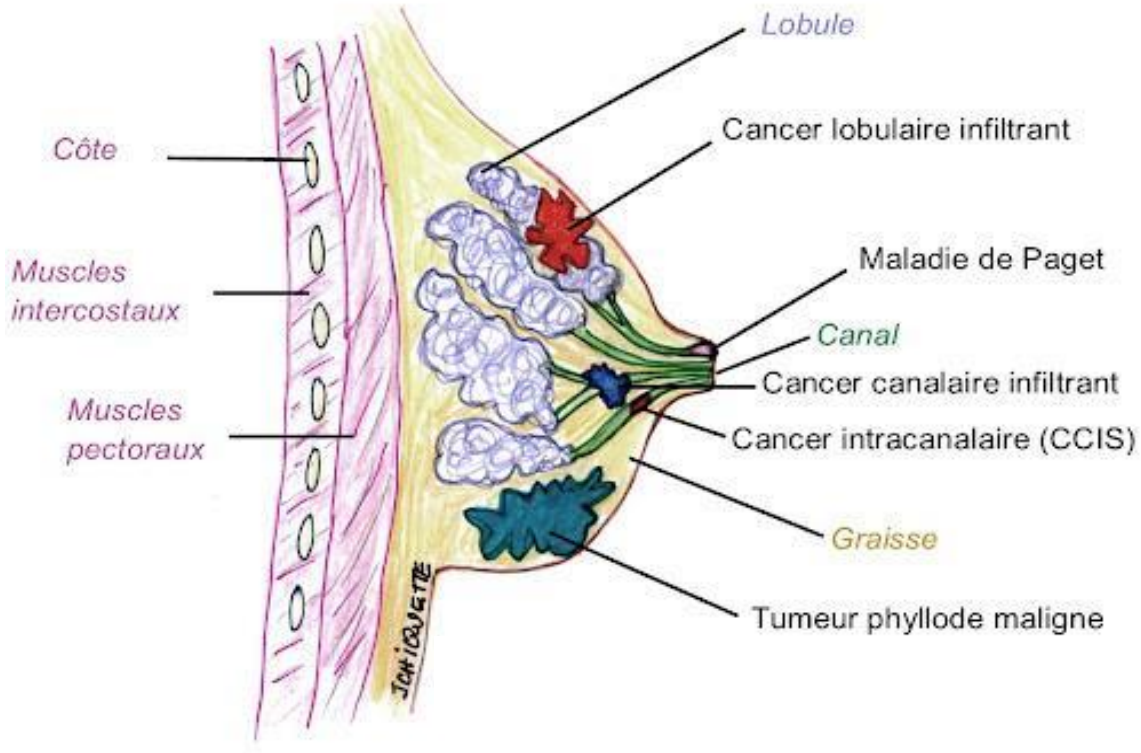


Figure I.6: Types de cancers des seins [15].

I.6.1. <u>Cancers non invasif :</u>

Carcinome Canalaire In Situ CCIS: est le type le plus fréquent de cancer du sein non invasif chez la femme. C'est une forme de cancer très «jeune», au tout début de sa formation. Comme son nom l'indique, il se forme a l'intérieure des canaux de lactation du sein. Il ne dissémine pas. On diagnostique beaucoup plus fréquemment ce type de cancer depuis l'utilisation plus répondre de la mammographie. Le traitement de ce cancer mène à la guérison dans presque tous les cas. Sans traitement, il poursuit sa croissance et peut alors devenir «infiltrant », donc se propager à l'extérieure des canaux galactophores [14].

Carcinome Lobulaire In Situ CLIS: les cellules cancéreuses apparaissent dans les lobules, puis traversent la paroi des lobules et se disséminent dans les tissus environnants [16].

I.6.2. Cancer invasif :

Carcinome Canalaire Infiltrant CCI: il se forme dans les canaux galactophores. Les cellules cancéreuses traversent la paroi des canaux.

Carcinome Lobulaire Infiltrant CLI: Carcinome inflammatoire forme dans les canaux mammaires puis migre vers les vaisseaux lymphatiques de la peau. Ce type de cancer progresse plus rapidement et est plus difficile à traiter [17], et est un cancer rare qui se caractériser principalement par un sein qui peut devenir rouge, enflé et chaud.

Autres carcinomes : médullaires, colloïdes ou muscinaux, tubulaires, et papillaires, où ces types de cancer du sein sont plus rares. Les principales différences entre ces types de cancer reposent sur le type de cellules touchées.

Maladie de Paget : Caractériser par une petite plaie au mamelon qui ne guérit pas se forme au niveau de la partie superficielle de la glande mammaire. Il est fréquemment associe à un autre cancer du sein [14,18].

I.7.Classification TNM des cancers du sein :

La classification TNM est un système de classement reposent sur l'extension tumorales locale, permet de décrire une tumeur, son possible envahissement ganglionnaire et l'éventuelle présence de métastases (Figure I.7). Il a été établir pour permettre des comparaisons en particulier internationales [19].

Son extension est décrite par trois paramètres T, N et M [20] :

T (tumeurs): correspond à a taille de la tumeur principale et à son degré d'extension. On classe T de x à 4 : un chiffre élevé signifie que la tumeur est grande ou qu'elle a envahi d'autres zones du sein ou des tissus proches.

N (node (ganglion)): représente le degré d'atteinte des ganglions lymphatiques situé à proximité de la tumeur.

M (métastase): est utilisé pour parler de l'extension du cancer du sein à d'autres parties du corps, et donc, la présence éventuelle de métastases.

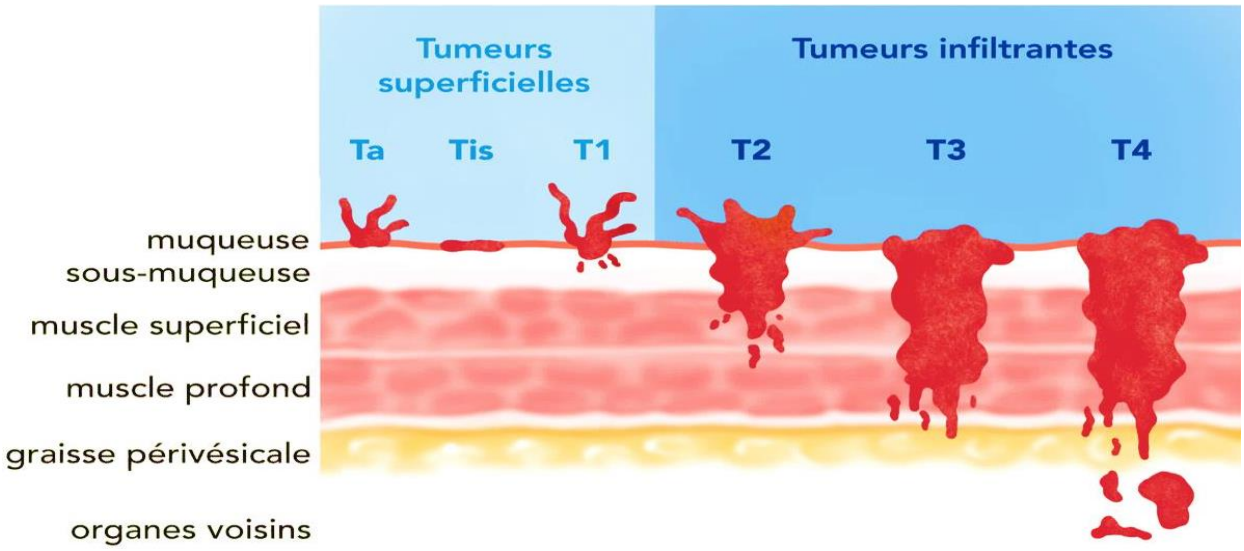


Figure I.7: les stades de classification de la tumeur de cancer (Tis est Carcinome in situ).

Ils sont classés dans le Tableau I.1

Tableau I.1: Classification TNM [21,22].

TNM	Stade	Localisation
T	X	Signifie que la tumeur est non évaluable.
	0	Tumeur non palpable.
	1	Tumeur inférieure ou égale à 2 cm.
	2	Tumeur supérieure à 2 cm et inférieure ou égale à 5 cm.
	3	Tumeur supérieure à 5 cm.
	4	a : tumeur étendue à la peau. b : à la paroi. c : aux deux. d : tumeur inflammatoires.
N	X	L'envahissement des ganglions lymphatiques ne peut pas être évalué.
	0	Absence de signe d'envahissement ganglionnaire axillaire.
	1	Tumeur inférieure ou égale à 2 cm.
	2	Adénopathies axillaires fixées entre elle ou aux parois de l'aisselle.
	3	Envahissement des ganglions mammaires internes.
M	0	Absence de métastases.
	1	Métastases (adénopathie sus-claviculaires incluses).

Pour les comparaisons, on peut regrouper les cas en stades selon le schéma habituel suivant [22]:

- Stade 0 : Tis N0 M0.

- Stade 1 : T1 N0 M0.
- Stade 2 : T1 N1 M0 et T2 N0 ou N1.
- Stade 3 : T1 N2, T2 N2, T3 N0 ou N1 ou N2.
- Stade 4 : T4 et/ou N3 et/ou M positif.

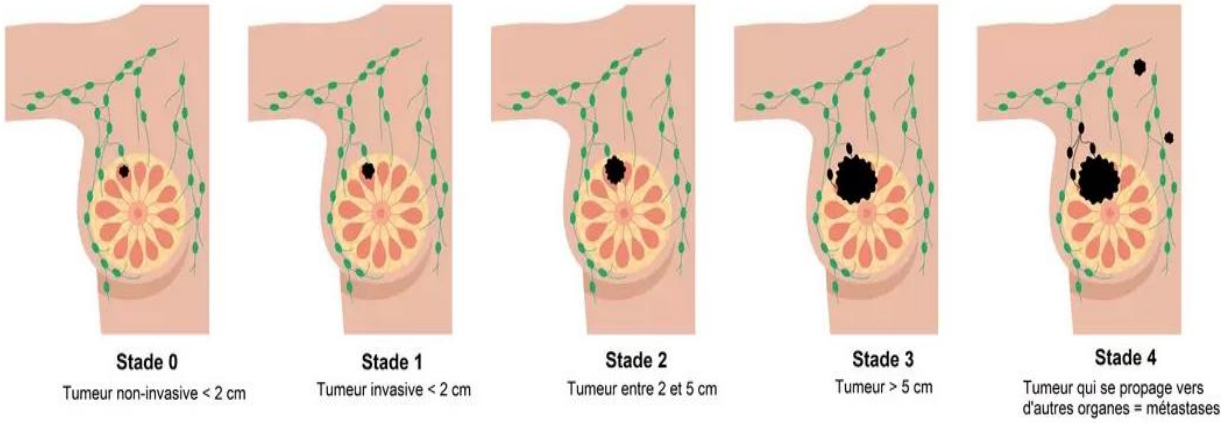


Figure I.8: Stades de développement du cancer du sein.

Les différents stades du cancer du sien décrits par la classification TNM sont illustres dans la Figure I.8 et les différentes caractéristiques de ces stades sont résumées dans le Tableau I.2.

Tableau I.2: les caractéristiques et statut migratoire des stades du cancer du sein.

STADE	CARACTÉRISTIQUES	STATUT MIGRATOIRE
0	Les cellules cancéreuses sont localisées seulement dans la membrane d'un canal galactophore (cancer dit carcinome canalaire in situ CCIS) ou dans la membrane d'un lobule (cancer dit carcinome lobulaire in situ CLIS) dans lequel elles ont pris naissance.	In situ (non-invasif) : Les cellules cancéreuses demeurent à l'endroit où elles ont pris naissances, sans envahir les tissus voisins.
1	La taille de la tumeur cancéreuse est de 2 centimètres ou moins et le cancer ne s'est pas propagé aux ganglions lymphatiques, ou on n'observe qu'un petit nombre de cellules cancéreuses dans les ganglions.	Infiltrant (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.
2	La taille de la tumeur est de plus de 2 cm et elle ne s'est pas propagée aux ganglions lymphatiques voisins, ou encore elle mesure moins de 5 cm,	Infiltrant (invasif) : la tumeur cancéreuse peut briser la membrane du tissu d'origine. Des cellules cancéreuses ont alors quitté leur tissu

	mais elle s'est propagée aux ganglions voisins.	d'origine et se sont localisées dans des tissus voisins.
3	Le cancer s'est propagé aux ganglions lymphatiques et possiblement aux tissus voisins, par exemple à des muscles ou à la peau.	Infiltrant (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.
4	Le cancer s'est propagé à d'autres parties du corps, dans des organes à distance.	Métastatique : Les cellules peuvent se mouvoir et emprunter la voie sanguine ou lymphatique pour se propager et se fixer sur d'autres organes.

I.8.Les facteurs de risque :

Un facteur de risque est quelque chose, comme un comportement, une substance ou un état, qui accroît le risque d'apparition d'un cancer. La plupart des cancers sont attribuables à de nombreux facteurs de risque, mais il arrive que le cancer du sein apparaisse chez des femmes qui ne présentent aucun des facteurs de risque décrits ci-dessous [23, 24].

I.8.1. Facteurs de risques non modifiables :

- L'âge : plus de 40 ans [25].
- La puberté précoce : premières règles avant l'âge de 12 ans.
- La ménopause tardive : après l'âge de 55 ans.
- Stérilité ou hypofertilité : pas ou peu d'enfant.
- Grossesse tardive : premier enfant après 30 ans [26].
- L'hérédité : cancer familiale [27].

I.8.2. Facteurs de risques modifiables :

- L'obésité : le rapport des graisses avec l'œstrogène.
- Le régime alimentaire : l'excès en aliment surtout les graisses d'origine animale.
- L'activité physique : l'activité physique diminuée surtout après l'âge de ménopause.
- L'allaitement : le non allaitement ou la durée courte d'allaitement.
- Tabagisme et alcool : les deux sont des facteurs de risques.
- La contraception orale et le traitement substitutif de la ménopause: (pilule).
- Les produits chimiques : les substances conservatrices, pesticides, déodorants.

- Le dépistage tardif du cancer du sein [28].

I.9. Le dépistage de cancer du sein :

Avant 50 ans, il est essentiel qu'une femme consulte chaque année son médecin pour qu'il puisse procéder à un examen clinique de ses seins. En cas de doute ou d'anomalie, il pourra alors programmer des examens complémentaires. Plusieurs outils existent pour établir un diagnostic de la pathologie [29].

I.9.1. La palpation : elle permet la mise en évidence d'une grosseur anormale.

Lorsqu'une personne présente des symptômes ou qu'une anomalie est décelée lors d'un examen de dépistage, un certain nombre d'examens doivent être réalisés afin d'établir un diagnostic. Toute suspicion diagnostic de cancer justifie un avis spécialisé sans délai. À plusieurs fins, notamment :

- Confirmer un diagnostic de cancer.
- Préciser le type histologique.
- Déterminer son ampleur (son stade) et son agressivité.
- Recueillir des prédicteurs connus de réponse à certains traitements.
- Identifier les contre-indications possibles à certains traitements.

Si le délai entre la détection d'une anomalie et le début du traitement semble parfois long, l'ensemble des tests réalisés dans le cadre de l'évaluation permettent de déterminer les options thérapeutiques les plus appropriées. Il ne faut jamais oublier que le cancer met plusieurs années à se développer [30].

I.9.2. L'imagerie :

- La mammographie : examen radiologique révélant des lésions de quelques millimètres indétectables par la palpation. Cette technique permet d'obtenir un cliché dans les 3 dimensions par reconstruction numérique ;
- L'échographie : examen utilisant les ultrasons, prescrit lorsque la mammographie a révélé une anomalie ou lorsque la densité des seins ne permet pas d'avoir une mammographie de qualité.

- L'imagerie par résonance magnétique (IRM) : elle est réalisée pour obtenir des renseignements complémentaires aux informations données par la mammographie et l'échographie.

I.9.3. Les prélèvements :

- La cytoponction : prélèvement de quelques cellules avec une aiguille très fine afin de les analyser.
- La biopsie : prélèvement d'un fragment de tissu réalisé sous anesthésie locale pour un examen microscopique [31].

I.10.Les Traitements :

Le choix de traitements dépend des caractéristiques suivantes [32]:

- Du type de cancer dont le patient atteint et de l'endroit où il est situé dans le sein.
- De son caractère unifocal (un foyer cancéreux) ou multifocal (plusieurs foyers cancéreux).
- De son stade au moment du diagnostic.
- De son grade.
- Du statut des récepteurs hormonaux ou de HER2.
- Des éventuelles contre-indications aux traitements.
- Santé général, de âge, de antécédents personnels médicaux et chirurgicaux et de antécédents familiaux du patient.
- De avis et préférences du patient.

Une proposition de traitements est établie par des médecins d'au moins trois spécialités différentes (chirurgien, oncologue médical, oncologue radiothérapeute, pathologiste...) dans le cadre d'une réunion de concertation pluridisciplinaire (RCP) en s'appuyant sur des recommandations de bonne pratique. La proposition de traitements est ensuite expliquée du patient au cours d'une consultation d'annonce. Lorsque vous avez donné votre accord sur cette proposition de traitement, ses modalités sont décrites dans un programme personnalisé de soins (PPS).

Si un cancer du sein est découvert, l'équipe soignante décide d'entamer un traitement, il existe plusieurs moyens pour traiter le cancer du sein [29] :

Chirurgie :

Les types de chirurgie qu'on le patient propose 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 :

La dosimétrie : la dose administrée est la quantité d'énergie distribuée dans les tissus par rayonnement. Cette énergie va entraîner des phénomènes qui aboutissent à la mort cellulaire. La dose se mesure et se calcule à l'aide de détecteurs ou de dosimètres [33].

Outre la dimension et l'orientation des faisceaux, l'étape de dosimétrie consiste à déterminer, par une étude informatisée, la distribution (autrement dit la répartition) de la dose de rayons à appliquer à la zone à traiter. Avec l'oncologue radiothérapeute, le physicien et le dosimétriste optimisent ainsi l'irradiation de façon à traiter au mieux la zone concernée tout en épargnant les tissus sains voisins [34].

Le plan de traitement définitif établit notamment la dose et ses modalités de délivrance (dose par séance, nombre et fréquence des séances …).

En cas de radiothérapie complémentaire d'une chirurgie conservatrice, réalisée à un niveau de dose totale importante de l'ordre de 46 à 70 Gy, délivrée sous forme de séances quotidiennes de 1,8 Gy à 2 Gy par séance sur une période de 5 à 7 semaines [35].

Dans certaines situations précises, la radiothérapie peut être administrée sur une plus courte période, pendant 3 semaines par exemple. C'est ce que l'on appelle un schéma dit hypofractionné.

Dans certains cas, une dose supplémentaire de 16 Gy peut être délivrée au niveau du lit tumoral en 1 à 2 semaines. Ce complément de dose (appelé boost ou surimpression) est parfois délivré par Curiethérapie [34].

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 [33].

I.11.Conclusion :

Le cancer de sein est une maladie très dangereuse pour le corps humain, et il résulte de la destruction de certaines cellules qui se développent souvent et forment une masse appelée "tumeur". Comme la détection précoce de cette maladie augmente le taux de guérison.

Récemment, de nombreuses études dans toutes disciplines ont contribué à fournir des méthodes. Prédire cette maladie avant l'infection repose sur l'intelligence artificielle et ses différentes branches.

Liste des références :

[1] Organisations mondiales de la santé. Site internet : https://www.who.int/fr/news-room/fact-sheets/detail/breast-cancer . [Consulté le 20/01/2024]

[2] Cancer-Comment apparaît le cancer ? (n.d.). Figaro Santé. https://sante.lefigaro.fr/sante/maladie/cancer-presentation-generale/comment-apparait-cancer. [Consulté le 13/01/2024]

[3] Bouaziz, H. Nouicer A, Boussouf N. Epidemiological and pathological profile of breast cancer in Southern Algerian women (2015-2019). https://doi.org/10.5281/zenodo.6025323. [Consulté le 13/01/2024]

[4] Corgne, A. 2016. Rôle du pharmacien d'officine dans la prise en charge du cancer du sein après chirurgie mammaire. Thèse de doctorat. Faculté de pharmacie de Dijon.p.20

[5] Anatomie du sein [Internet]. Dr. Anne Wautier, gynécologue. Disponible sur: https://sante.journaldesfemmes.fr/fiches-anatomie-et-examens/2571039-sein-anatomie-examens-et-maladies/. [Consulté le 15/01/2024]

[6] Sante.journaldesfemmes.fr. https://sante.journaldesfemmes.fr/fiches-anatomie-et-examens/2571039-sein-anatomie-examens-et-maladies/. [Consulté le 20/01/2024]

[7] Lee, S. (2012). Les seins. Société Canadienne Du Cancer. https://cancer.ca/fr/cancer-information/cancer-types/breast/what-is-breast-cancer/the-breasts. [Consulté le 22/01/2024]

[8] Runbin, P. Hansen, JT. TNM Staging Atlas with Oncoanatomy. 2nd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2012: http://www.lwwoncology.com. [Consulté le 20/01/2024]

[9] l'ISHH, L. de. (2022, October 2). Symptômes Cancer du sein, signes avant-coureurs & douleurs aisselles. ISHH. https://ishh.fr/cancer-du-sein/symptome-signes-avant-coureurs/. [Consulté le 20/02/2024]

[10] Saunders, C. Jassal, S. (2009). Breast cancer (1. ed.). Oxford: Oxford University Press. p. Chapter 13. ISBN 978-0-19-955869-8. Archived from the original on 25 October 2015.

[11] Molckovsky, A. Fitzgerald, B. Freedman, O. Heisey, R. & Clemons, M. (2009). Approach to inflammatory breast cancer. Canadian Family Physician, 55(1), 25-31.

[12] "Breast Disorders: Breast Cancer". Merck Manual of Diagnosis and Therapy. February 2003. Archived from the original on 2 October 2011. Retrieved 5 February 2008.

[13] American Society of Clinical Oncology. Breast Cancer: Introduction (https://www.cancer.net/cancer-types/breast-cancer/introduction). Multiple pages reviewed. Last updated 10/2022. [Consulté le 22/01/2024]

[14] LORIOT, Y. MORDAUT, P. (2011) « cancérologie », édition Elsevier Masson.

[15] Carcinome infiltrant. (n.d.). Www.depistagesein.ca. Retrieved March 23, 2024, from http://www.depistagesein.ca/carcinome-infiltrant/. [Consulté le 24/01/2024]

[16] Bodian, CA. Espié, M. Page, DL. Roquancourt, A.(2014).Prise en charge thérapeutique du carcinome lobulaire in situ .Article paru dans le Genesis.,N°180

[17] Bergaoui, H. El Mhabrech, H. Zouari, I. Njima, M. Daldoul, A. Ahmed, H. Faleh, R. (2019). Le carcinome lobulaire infiltrant du sein : à propos de 30 cas. Pan African Medical Journal, 34(70)

[18] HAS. Actualisation du référentiel de pratiques de l'examen périodique de santé - Dépistage et prévention du cancer du sein. Février 2015. Disponible sur le site de la HAS. Consulté le 28 juin 2019

[19] Cristina, DA. Marc, DB. Ahmed, DB. Anca, DB. Cécile, DBF. Françoise, DC-L. et al. Référentiel Cancer du sein invasif- Onco Normandie. 2018.

[20] Classification du cancer : TNM, grade, stade du cancer | FQC. (n.d.). Retrieved March 19, 2024, from https://cancerquebec.ca/information-sur-le-cancer/le-cancer/classification-cancer/. [Consulté le 20/02/2024]

[21] Morrow, M. Burstein, HJ. and Harris, JR. Malignant tumors of the breast. Devita, VT, Jr. Lawrence, TS. Rosenberg, SA. Cancer: Principles and Practice of Oncology. 10th ed. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2015: 79: 1117-1156.

[22] Hortobagyi, GN. Connolly, JL. D'Orsi, CJ. et Breast, Al, Amin, MB. (ed.). AJCC Cancer Staging Manual. 8th ed. Chicago, IL: American College of Surgeons; 2017: 48:589–636.

[23] Institue internationale du cancer situé : https://www.e-cancer.fr/Patients-et-proches/Les-cancers/Cancer-du-sein/Facteurs-de-risque-de-recidive. [Consulté le 08/03/2024]

[24] Fakhri, N. Chad, MA. Lahkim, M. Houari, A. Dehbi, H. Belmouden, A. El Kadmiri, N. (September 2022). "Risk factors for breast cancer in women: an update review". Medical Oncology. 39 (12): 197.

[25] "Breast Cancer Treatment (PDQ®)". NCI. 26 June 2014. Archived from the original on 5 July 2014. Retrieved 29 June 2014

[26] Le Médecin du Québec, « SEINvestir dans la prévention », volume 45, numéro 10,oct. 2010. https://lemedecinduquebec.org/Media/108283/051-056DresGauthier_Dostie.pdf . [Consulté le 08/03/2024]

[27]Masson, E. « Épidémiologie du cancer du sein », EM-Consulte. https://www.emconsulte.com/article/711266/epidemiologie-du-cancer-du-sein.[Consulté le 24/01/2024]

[28] Facteurs de risque. (n.d.). Institut de Cardiologie de Montréal. https://www.icm-mhi.org/fr/prevention/adopter-de-saines-habitudes-de-vie/facteurs-de-risque.[Consulté le 24/01/2024]

[29] Cancer Society. (2021). Breast Cancer Treatment. Retrieved from. https://www.cancer.org/cancer/breast-cancer/treatment.html. [Consulté le 20/01/2024]

[30] Le cancer du sein : Cancer du sein. (n.d.). Www.e-Cancer.fr.https://www.e-cancer.fr/Patients-et-proches/Les-cancers/Cancer-du-sein/Les-points-cles.

[Consulté le01/05/2024]

[31]Breast.cancer.org.Metastatic.Breast.Cancer
https://www.breastcancer.org/types/metastatic#section-what-is-metastatic-breast-cancer.
[Consulté le 22/03/2024]

[32] Guide Cancer Info « Les traitements des cancers du sein »- novembre 2013

[33] ARCAGY-GINECO, D. B. P. -. (2024, May 7). Les doses utilisées en radiothérapie. Infocancer.https://www.arcagy.org/infocancer/traitement-du-cancer/traitements-locoregionaux/radiotherapie/les-doses-de-radioactivite.html/. [Consulté le 01/05/2024]

[34] Du, N. (2019). Radiothérapie externe - Radiothérapie. E-Cancer.fr. https://www.e-cancer.fr/Patients-et-proches/Les-cancers/Cancer-du-sein/Radiotherapie/Radiotherapie-externe. [Consulté le 01/05/2024]

[35] FAQ : la radiothérapie | Centre de Radiothérapie de Bobigny - IRHE. (2024). Ramsaysante.fr.https://irhe-bobigny.ramsaysante.fr/faq-la-radioth%C3%A9rapie. [Consulté le 01/05/2024]

Chapitre II :

Apprentissage automatique

II.1.Introduction :

Le développement de la technologie continue à prendre une place de plus en plus importante dans le domaine de la santé, ce qui a permis le développement du matériel médical, des logiciels de surveillance médicale et des logiciels d'analyses médicales qui augmentent la précision des résultats.

L'intelligence artificielle est l'un des domaines les plus actifs de nos jours en informatique et elle est représenté l'avenir de l'informatique classique [1]. Les principaux sous-domaines de l'IA comprennent l'apprentissage automatique (machine learning), le traitement du langage naturel, les systèmes experts, la vision par ordinateur et la robotique. Il est important de mentionner que ces sous-domaines ne sont pas exclusifs ; ils se chevauchent souvent. Selon Eric Sibony [2], « machine » ne désigne pas un objet physique, mais plutôt « un système automatique capable de traiter de data et de l'information ».

Les technologies intègrent désormais l'intelligence artificielle (IA) dans le processus thérapeutique; les outils d'IA mettent en œuvre des algorithmes dans les machines à rayonnement et les équipements d'imagerie pour administrer avec précision une radiothérapie [3]. Le besoin de l'IA dans la pratique de la radiothérapie est devenu nécessaire associé à l'incidence croissante du cancer [4,5].

Dans ce chapitre, nous exposons tout d'abord la définition de la machine learning. Ensuite, nous présentons les concepts et les techniques les plus importantes utilisées.

II.2.Définition de l'apprentissage automatique :

L'apprentissage automatique (machine learning en anglais) est une forme d'intelligence artificielle qui vise à donner aux machines la capacité «d'apprendre » à partir de données, via des modèles mathématiques [6**].** Contrairement à la programmation traditionnelle où l'humain doit écrire un programme et fournir à l'ordinateur des entrées pour avoir des résultats en sortie, un algorithme d'apprentissage machine prend des entrées ainsi que les résultats voulus et génère un programme (une fonction mathématique) en sortie [7].

Une fois l'apprentissage réalisé et le modèle validé, le modèle pourra ensuite être déployé en production pour une utilisation réelle [8].

II.3.Types d'apprentissage automatique :

L'apprentissage automatique a plusieurs approches qui permettent d'extraire les caractéristiques cachées dans les données. On peut classer ces techniques en 2 grandes catégories Ils sont résumés dans la figure II.1 :

- Apprentissage supervisé.
- Apprentissage non supervisé.

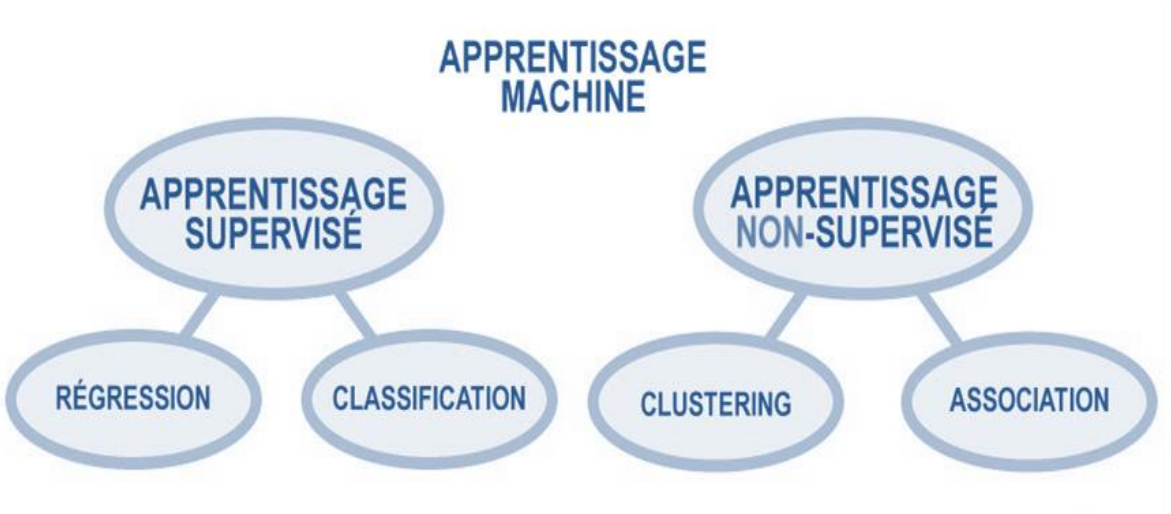


Figure II.1: taxonomie des différentes techniques issues de l'apprentissage automatique.

La différence entre les deux types en résumé dans le Tableau II.1.

Tableau II.1: Différences entre l'apprentissage supervisé et non supervisé.

Apprentissage supervisé	Apprentissage non supervisé
Données d'entrée sont étiquetées.	Données d'entrée son non étiquetées.
Utilise le jeu de données d'apprentissage.	Utilise tout le jeu de données en entrée.
Utilisé pour la prédiction.	Utilisé pour l'analyse.
Classification et régression.	Regroupement, estimation de la densité, et réduction de la dimensionnalité.

II.3.1.Apprentissage automatique non supervisé :

Dans l'apprentissage non supervisé, le réseau est fourni avec des entrées, mais pas avec les sorties souhaitées. Le système lui-même doit alors décider quelles fonctionnalités il utilisera pour regrouper les données d'entrée. C'est ce qu'on appelle souvent l'auto-organisation ou l'adaptation [9, 10]. Les étapes d'apprentissage non supervisé peuvent être représentées par la figure II.2.

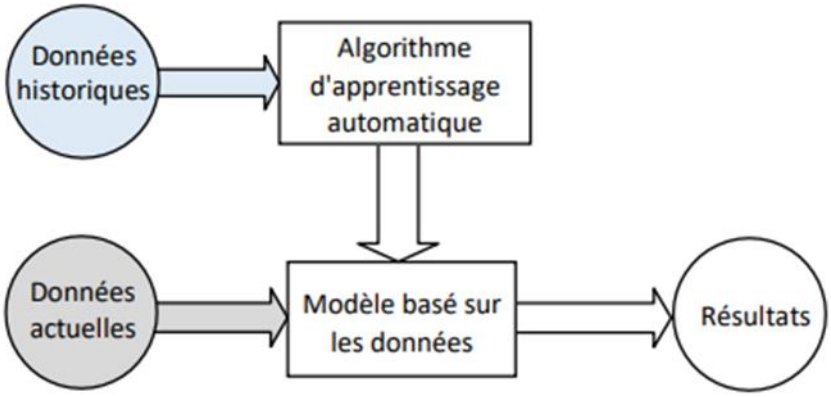


Figure II.2: apprentissage automatique non supervisé.

L'apprentissage automatique non supervisé peut être comparé à un enfant qui apprend à identifier le type de fruit en observant le motif et la couleur, au lieu de mémoriser les noms avec l'aide d'une autre personne. Il recherche des similitudes entre les images, les séparant ainsi en groupes, tout en attribuant à chaque groupe son propre label [11].

a. **Clustering (Regroupement) :** c'est une méthode d'analyse statistique qui vise à collecter des données en fonction de leurs similarités ou de leurs différences, utilisée pour organiser des données brutes en silos homogènes à l'intérieur de chaque grappe. Les algorithmes les plus célèbres utilisés dans cette approche sont [12]:

K-means (K-Moyenne) : est un algorithme d'apprentissage non supervisé, est un algorithme de regroupement (clustering), qui regroupe l'ensemble des données non étiquetées en différentes grappes [13].

Il s'agit d'un algorithme itératif qui divise l'ensemble de données non étiqueté en k clusters différents de telle sorte que chaque ensemble de données n'appartienne qu'à un seul groupe ayant des propriétés similaires, sans aucune formation.

Il s'agit d'un algorithme basé sur le centroïde, dans lequel chaque cluster est associé à un centroïde. L'objectif principal de cet algorithme est de minimiser la somme des distances entre les points de données et leurs clusters correspondants.

L'algorithme de clustering k-means effectue principalement deux tâches :

- Détermine la meilleure valeur pour K points centraux ou centroïdes par un processus itératif.
- Attribue chaque point de données à son centre K le plus proche. Les points de données qui sont proches du centre k particulier créent un cluster.

Par conséquent, chaque cluster possède des points de données présentant certains points communs et est éloigné des autres clusters.

Hierarchical clustering HCA (Analyse de classification hiérarchique) : La création d'un cluster hiérarchique ressemble à la création d'un cluster normal, à l'exception que dans cette situation, nous souhaitons instaurer une hiérarchie des clusters. Cela peut être extrêmement crucial, en particulier lorsque nous souhaitons avoir une grande souplesse quant au nombre de clusters souhaités.

b. Association :

Apriori : Dans une base de données transactionnelle, l'algorithme Apriori est employé afin d'extraire des ensembles d'éléments fréquents, puis de générer des règles d'association.

II.3.2.Apprentissage automatique supervisés :

L'apprentissage supervisé est un type d'apprentissage automatique qui utilise un ensemble de données connu pour effectuer des prédictions. L'ensemble des données d'apprentissage se compose de données d'entrée et de valeurs de réponse. Les algorithmes d'apprentissage supervisé cherchent à créer un modèle capable de prédire les valeurs de réponse d'un nouvel ensemble de données [14, 15]. Les étapes d'apprentissage supervisé peuvent être représentées par la figure II.3.

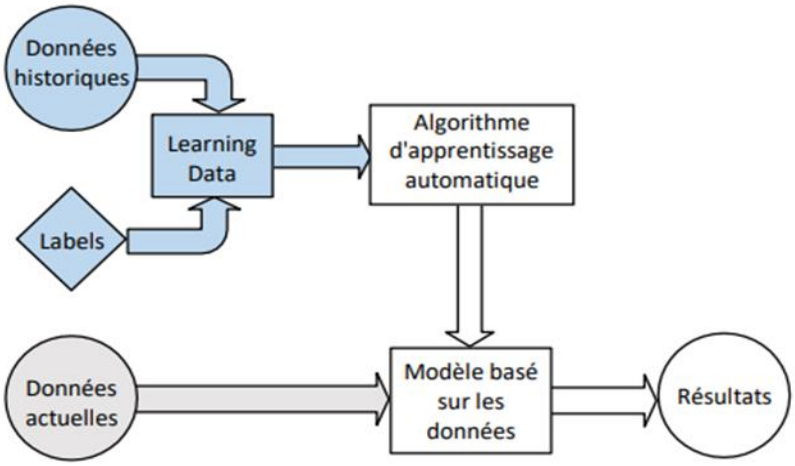


Figure II.3: Apprentissage automatique supervisé.

Le processus d'apprentissage supervisé se déroule généralement en plusieurs étapes :

1. Entraînement : Pendant cette étape, l'algorithme est exposé à un ensemble de données d'entraînement qui contient des exemples étiquetés. L'algorithme ajuste ses paramètres internes en analysant ces données pour apprendre à faire des prédictions précises.
2. Validation : Après l'entraînement, l'algorithme est évalué sur un ensemble des données de validation distinct pour estimer ses performances et ajuster ses hyper-paramètres (paramètres qui contrôlent le processus d'apprentissage).
3. Test : Une fois que l'algorithme a été entraîné et validé, il est testé sur un ensemble de données de test distinct pour évaluer ses performances réelles sur des données qu'il n'a jamais vues auparavant. Cela permet d'estimer à quel point l'algorithme est précis par rapport aux données non vues [16].

L'apprentissage supervisé consiste en des variables d'entrée (X) et une variable de sortie (Y). Vous utilisez un algorithme pour apprendre la fonction de l'entrée à la sortie.

$$Y = f(X) \qquad \textbf{(II.1)}$$

Le but est d'appréhender si bien la fonction que, lorsque vous avez de nouvelles données d'entrée (X), vous pouvez prédire les variables de sortie (Y) pour ces données. Nous connaissons les réponses correctes, l'algorithme effectue des prédictions itératives sur les données d'apprentissage et est corrigé. L'apprentissage s'arrête lorsque l'algorithme atteint un niveau de performance acceptable. C'est ce qu'on appelle l'apprentissage supervisé.

Les algorithmes de l'apprentissage automatique supervisé sont les plus couramment utilisés, et il y a deux types d'apprentissage supervisé :

a. **La classification:** la classification consiste à trouver le lien entre une variable d'entrée (X) et une variable de sortie discrète (Y), en suivant une loi multinomiale [17]. Voici quelques exemples populaires d'algorithmes de classification [18]:

Support Vector Machine SVM (Machine à vecteurs de support): Il s'agit d'un modèle très puissant et polyvalent d'apprentissage automatique, capable de réaliser la classification linéaire ou non linéaire. Cette approche de l'apprentissage automatique « Machine Learning » est l'un des modèles les plus appréciés. Les SVM se révèlent particulièrement performants pour classifier des ensembles de données complexes mais de taille réduite à moyenne. L'objectif de l'algorithme SVM est de trouver à la fois le plan hyperoptimal et de réduire les erreurs de classification et trouver le meilleur séparateur (ligne, plan ou hyperplan) qui sépare le mieux les deux types (figure II.4).

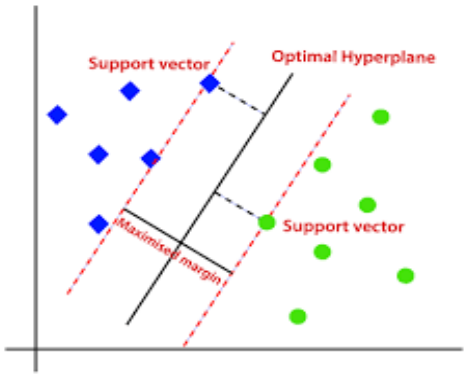


Figure II.4: Principe de Classification en SVM.

K-Nearest Neighbors KNN (K-plus proches voisins): cet algorithme consiste à essayer différente valeurs de K pour obtenir la séparation la plus satisfaisante.

Par exemple Dans la figure II.5, si on choisit k = 3, l'algorithme cherche les trois plus proches voisins du cercle rouge pour pouvoir le classé soit dans la classe des cercles, soit dans la classe des carrés.

Dans ce cas, les trois plus proches voisins du cercle rouge sont un carré et deux cercles. Par conséquent, l'algorithme classera le cercle rouge dans la classe des cercles.

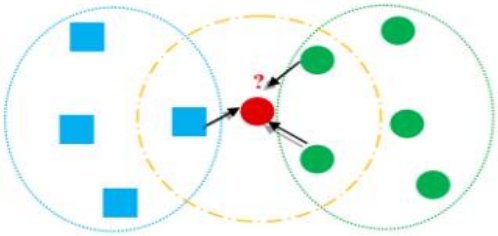


Figure II.5: Exemple de classification par l'algorithme KNN.

Gaussian Naïve Bayes (La classification naïve bayésienne): Est la méthode de classification qui classe un ensemble d'observations selon les règles déterminées par l'algorithme lui-même. Il s'appuie sur le théorème de Bayes des probabilités conditionnelles et suppose que les variables sont autonomes les unes des autres. Cela facilite la détermination des probabilités. L'outil de classification doit d'abord être entraîné sur un ensemble de données d'entraînement, qui affiche la catégorie attendue en fonction de l'entrée. Dans la phase d'apprentissage, l'algorithme développe ses règles de classification sur cet ensemble de données et les applique à la classification de l'ensemble de données prédit dans la deuxième étape. Le classificateur NB signifie que la classe des données d'apprentissage est connue et fournie.

Decision tree (Les arbres de décision) : L'arbre de décision peut être construit par une approche algorithmique qui peut diviser l'ensemble de données de différentes manières en fonction de différentes conditions [19]. Tel qu'il est représentant un ensemble de choix sous la forme graphique d'un arbre. Les différentes décisions possibles sont situées aux extrémités des branches (les « feuilles » de l'arbre), et sont atteintes en fonction de décisions prises à chaque étape.

Logistic regression (Régression logistique): Est destiné aux tâches de classification. L'algorithme de régression logistique vise à identifier les coefficients les plus appropriés afin de réduire l'erreur entre la prédiction effectuée pour des destinations visitées et la vraie étiquette donnée (par exemple, bon, mauvais, etc.).

Random Forest (Forêt aléatoire): sont une approche globale pour la classification, la régression et d'autres tâches, qui fonctionnent en créant de nombreux arbres de décision lors de l'entraînement et en générant la classe qui représente le mode de classification ou la prédiction moyenne (régression) des arbres individuels. Aléatoire les forêts de décision corrigent l'habitude des arbres de décision de trop s'adapter à leur ensemble d'apprentissage.

b. Régression : la régression consiste à prédire une valeur continue pour la variable de sortie [20], les algorithmes de régression tentent d'estimer la fonction (f) des variables d'entrée (x) aux variables de sortie numériques ou continues (y). Il existe plusieurs modèles pour la régression [18] :

Linear regression (Régression linéaire) : La régression linéaire multiple a comme but de décrire la variation d'une variable dépendante (y) associée aux variations de plusieurs variables indépendantes. Dans le contexte de l'apprentissage automatique, elle sert à estimer une fonction linéaire entre la sortie (avec des valeurs continues, numériques) et les entrées. La fonction qui estimé les valeurs de y d'un échantillon en se basant sur des caractéristiques d'entrée x est écrite comme suit :

$$y'(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_m x_m \qquad \textbf{(II.2)}$$

Dans lequel :

- y' : variable dépendante (variable cible).
- x : variable indépendante (variable attendue).
- θ_0 : interception de ligne (donne un degré de liberté supplémentaire).
- θ_1 : Coefficient de régression linéaire (facteur d'échelle pour chaque valeur d'entrée).

Les valeurs des variables x et y sont des ensembles de données d'entraînement pour représenter un modèle de régression linéaire.

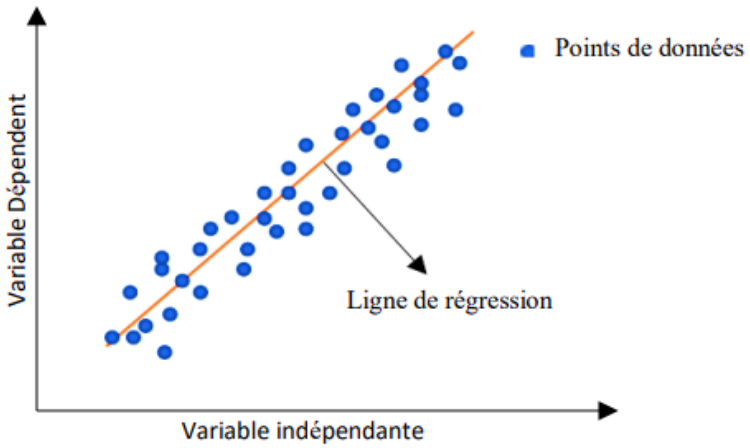


Figure II.6: Un modèle de régression linéaire.

Support Vector Regression (SVR) (Régression à vecteur de support): Régression à vecteur de support SVR cherche à trouver une hyperplane qui correspond le mieux aux points de données dans un espace continu. Ceci est obtenu en cartographiant les variables d'entrée dans un espace de caractéristique haute dimension et en trouvant l'hyperplane qui maximise la marge (distance) entre l'Hyperplan et les points de données les plus proches, tout en minimisant également l'erreur de prédiction.

SVR peut gérer les relations non linéaires entre les variables d'entrée et la variable cible en utilisant une fonction de noyau pour cartographier les données dans un espace plus élevé. Cela en fait un outil puissant pour les tâches de régression où il peut y avoir des relations complexes entre les variables d'entrée et la variable cible [21].

- ✓ **Kernel (noyau):** La fonction de conversion d'un ensemble de données de dimension inférieure en un ensemble de données de dimension supérieure, un noyau aide à la recherche d'un hyperplan dans un espace de dimension supérieure tout en réduisant le coût de calcul. Le choix de la fonction du noyau affecte les performances et la flexibilité du modèle régression du vecteur de support SVR. En suit quelques fonctions de noyau couramment utilisées pour SVR :

  1. Linear kernel : Le noyau linéaire est la fonction noyau la plus simple et fonctionne bien lorsque la relation entre les variables d'entrée et de sortie doit être linéaire. Il est défini comme [22]:

  $$K(x, x') = x * x' \quad \textbf{(II.3)}$$

  2. Polynomial kernel : La fonction de noyau polynomial introduit la non-linéarité en transformant les caractéristiques d'origine en un espace de dimension supérieure à l'aide de fonctions polynomiales. Il est défini comme [22]:

  $$K(x, y) = (a * x * y + b)^d \quad \textbf{(II.4)}$$

  Ici, **a** et **b** sont des paramètres définis par l'utilisateur qui contrôlent la forme du polynôme, et **d** spécifie le degré du polynôme.

  3. Radial Basis Function (RBF) Kernel : Le noyau RBF est un choix populaire pour SVR en raison de sa flexibilité dans la capture de relations complexes. Il mappe les données d'entrée dans un espace de dimension infinie à l'aide de fonctions gaussiennes. Le noyau RBF est défini comme suit [22]:

  $$K(x, y) = \exp(-a * ||x - x'||^2) \quad \textbf{(II.5)}$$

Ici, **a** est un paramètre défini par l'utilisateur qui détermine l'influence de chaque exemple de formation. Des valeurs plus élevées de a conduisent à une frontière de décision plus localisée et ondulée.

4. **Sigmoid Kernel :** Le noyau sigmoïde est un autre noyau non linéaire qui est utile lorsqu'il s'agit de relations non linéaires. Il est défini comme [23]:

$$K(x, y) = \tanh(a * x * y + b) \quad \textbf{(II.6)}$$

Les paramètres **a** et **b** contrôlent la forme et la non-linéarité de la fonction sigmoïde.

II.4.L'intelligence artificielle et le traitement de cancer :

La science s'applique depuis longtemps à élaborer des méthodes de dépistage et de diagnostic permettant de mieux identifier et catégoriser les tumeurs mammaires.

Les algorithmes de ML qui jouent un rôle de plus en plus important quant à la prédiction d'un probable développement de la maladie chez différents patients [24]. En effet, le taux de précision de la prédiction du cancer a augmenté de 15 à 20 % sur les dernières années grâce à l'utilisation du ML [20].

Une intelligence artificielle est désormais capable de prédire une grande probabilité de risque d'avoir un cancer. Malgré toutes ces techniques à disposition, la plupart des modèles proposés présentent toujours des limites. Les efforts devront toutefois se poursuivre afin de soutenir le passage de l'IA du stade de la recherche à celui de la pratique.

II.5.Conclusion :

En conclusion, les technologies d'intelligence artificielle présentent d'importantes opportunités pour faire progresser le traitement du cancer. Grâce à sa capacité à analyser et à traiter les données, l'IA peut aider les professionnels de la santé à prendre des décisions plus rapides et plus précises, à personnaliser les traitements en fonction des caractéristiques uniques de chaque patient et à découvrir des traitements nouveaux et prometteurs. Même si des problèmes subsistent, notamment en termes de qualité des données et d'acceptation clinique, l'IA représente un outil potentiellement révolutionnaire dans la lutte contre le cancer.

Liste de références :

[1] Kazar,Okba. Intelligence artificielle et ses applications. Last accessed 02 July 2019

[2] «Au fait, c'est quoi l'intelligence artificielle ?». Available:https://www.europe1.fr/technologies/au-fait-cest-quoi-lintelligence-artificielle-3612572. [Consulté le 23/04/2024]

[3] Malicki, J. Medical physics in radiotherapy: The importance of preserving clinical responsibilities and expanding the profession's role in research, education, and quality control. Rep Pract Oncol Radiother 2015;20:161-9.

[4] Jiang, L. Wu, Z. Xu, X. et al. Opportunities and challenges of artificial intelligence in the medical field: current application, emerging problems, and problem-solving strategies. J Int Med Res 2021;49:3000605211000157.

[5] El Naqa, I. Haider, MA. Giger, ML. et al. Artificial Intelligence: reshaping the practice of radiological sciences in the 21st century. Br J Radiol 2020;93:20190855.

[6] Géron, A. (2017). Hands-On_Machine_Learning_with_Scikit-learn and trensflow. USA: O'Reilly Media

[7] L'apprentissage automatique : un atout puissant pour une meilleure exploitation de vos données. (n.d.). Bba.https://www.bba.ca/ca-fr/publications/lapprentissage-automatique-un-atout-puissant-pour-une-meilleure-exploitation-de-vos-donnes. [Consulté le 22/04/2024]

[8] Apprentissage automatique. (2024). Cnil.fr. https://www.cnil.fr/fr/definition/apprentissage-automatique. [Consulté le 25/04/2024]

[9] Gurney, K. an introduction to neural networks. 1997.

[10] Fausett, Lv. fundamentals of neural networks: architectures, algorithms and applications. 1993.

[11] Hilali, H. application de la classification textuelle pour l'extraction des règles d'association maximales. Thèse de maitrise en informatique, université du québec à trois-rivières, trois-rivières, 2009.

[12] Issarane, H. (2019, 02 09). Apprentissage Non Supervisé. Frome Le DataScientist: https://le-datascientist.fr/apprentissage-non-supervise. [Consulté le 23/04/2024].

[13] Javatpoint. (n.d.). K-Means Clustering Algorithm - Javatpoint. Www.javatpoint.com. https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning. [Consulté le 23/04/2024]

[14] NF EN 13306, Norme européenne, Éditée et diffusée par l'Association Française de Normalisation (AFNOR), Juin 2001.

[15] Patrick, Jahnke. «Machine Learning Approaches for Failure Type Detection and Predictive Maintenance», June 19, 2015.

[16] Apprentissage supervisé vs non-supervisé en Data Science. (n.d.), from https://www.data-bird.co/blog/apprentissage-supervise-vs-non-supervise#:~:text=Contrairement%20%C3%A0%20l. [Consulté le 25/04/2024]

[17] Dupré, X. (2020). La classification. Retrieved 19 10,2021 from Xavierdupre: http://www.xavierdupre.fr/app/mlstatpy/helpsphinx/c_ml/rn/rn_ 3_clas.htm. [Consulté le 22/01/2024]

[18] GAËL. (2019, 10 25). Machine Learning. from Datakeen: https://datakeen.co/8-machine-learning-algorithms-explained-in-human-language/. [Consulté le 23/04/2024]

[19] Contributeurs aux projets Wikimedia. (2005, September 14). Outil d'aide à la décision. From: https://fr.wikipedia.org/wiki/Arbre_de_d%C3%A9cision. [Consulté le 22/05/2024]

[20] Kourou, Konstantina. Et al. "Machine learning applications in cancer prognosis and prediction." Computational and structural biotechnology journal 13 (2015): 8-17.

[21] Sethi, A. (2020, March 27). *Support Vector Regression In Machine Learning*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/.[Consulté le 24/05/2024]

[22] C, H. Wu, G H. Tzeng et R H. Lin, « A novel hybrid genetic algorithm for kernel function and parameter optimization in support vector regression, » Expert Systems with Applications, t. 36, no 3, p. 4725-4735, 2009 (cf. p. 33).

[23] M S, Ahmed et al, « Yield Response of Different Rice Ecotypes to Meteorological, Agro-Chemical, and Soil Physiographic Factors for Interpretable Precision Agriculture Using Extreme Gradient Boosting and Support Vector Regression., » Complexity, 2022 (cf. p. 33, 34).

[24] Intelligence artificielle : définition, histoire et application. (n.d.). Www.justai.co. Retrieved April 17, 2024, from https://www.justai.co/articles-de-blog/intelligence-artificielle. [Consulté 22/04/2024]

Chapitre III :

Résultats et discussion

III.1.Introduction :

La première section de ce chapitre présente une étude statistique des données des patients. Puis, nous exposons les résultats obtenus à travers l'implémentation des différents algorithmes d'apprentissage (régression, classification binaires). Nous effectuons également une comparaison entre les performances des modèles générés. Le reste du chapitre consiste à tester la capacité prédictive de nos modèles sur nos jeux de données en utilisant les métriques et nous comparons nos résultats.

III.2.Partie appliquée :

Notre étude a réalisé dans le Centre Anti Cancéreux (CAC) de l'EPH Mohamed Boudiaf Ouargla.

L'autorité a été contactée pour obtenir l'accord préalable du service administratif afin de réaliser une étude statistique des patientes atteintes d'un cancer du sein. Il s'agit d'une étude portant sur les patientes ayant été prise en charge pour de CS au niveau du CAC – Ouargla- durant la période allant du 2018 - 2023.

Cette tâche a demandé beaucoup de temps et d'efforts, car tous les dossiers médicaux étaient manuscrits et beaucoup étaient rédigés en espagnol. Finalement, nous avons pu identifier environ 600 dossiers, dont 458 ont été utilisés contenant les données nécessaires à cette étude.

Le CAC d'Ouargla est situé au sein de l'Etablissement public hospitalier (EPH) Mohammed Boudiaf de Ouargla.

Il a été mis en service en 2009 avec une capacité d'hospitalisation de 84 lits réparties en quatre unités :

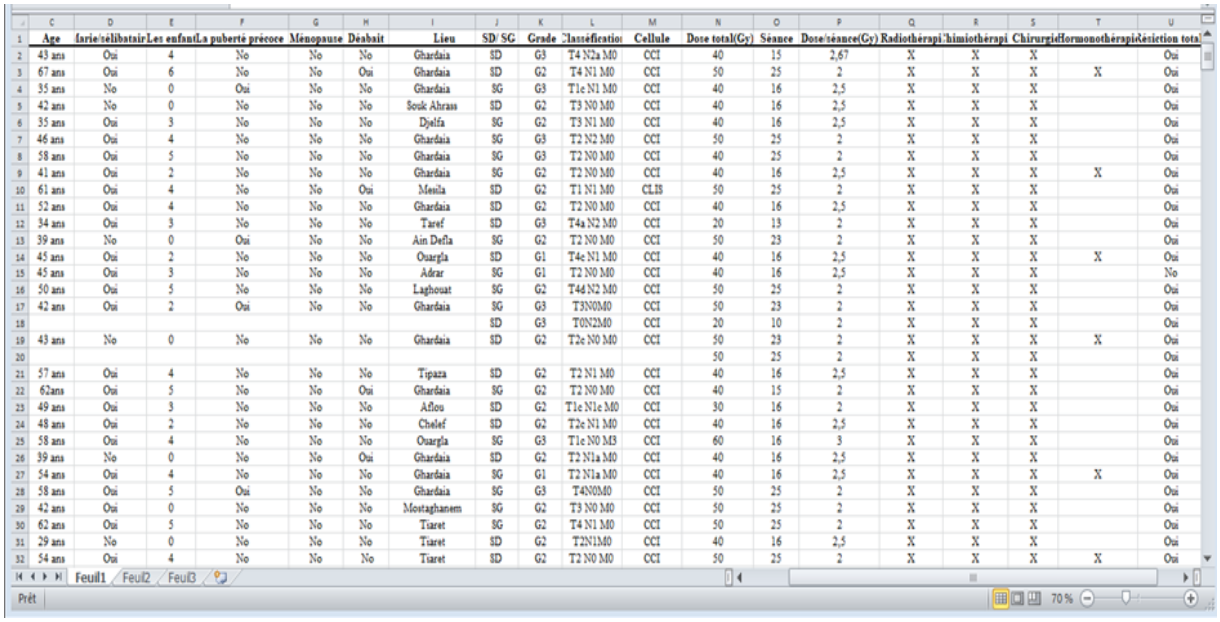
- Chirurgie oncologique.
- Oncologie médicale.
- Radiothérapie.
- Unité de médecine nucléaire.

C'est le premier CAC du sud. Jusqu'en 2018 il couvrait toute la région du Sud algérien (tel que : Ouargla, El oued, Ghardaïa, Djelfa, Illizi, Tamanrasset, Laghouat et Adrar).

Le CAC dispose de deux appareils de haute énergie (télé cobalt et accélérateur) et un appareil de curiethérapie utilisés en radiothérapie, deux appareils caméra-gamma et un labo chaud utilisés en médecine nucléaire.

III.2.1. Recueil des données :

Les données relatives à chaque sujet de l'étude ont été obtenues à partir des dossiers médicaux des patientes à partir d'archive de service de radiothérapie. Ces informations ont été répertoriées sur le fiche indiqué dans la figure III.1.



Age	Marie/célibataire	Les enfants	La puberté précoce	Ménopause	Déabait	Lieu	SD/ SG	Grade	Classification	Cellule	Dose total(Gy)	Séance	Dose/séance(Gy)	Radiothérapie	Chimiothérapie	Chirurgie	Hormonothérapie	Résiction tota
43 ans	Oui	4	No	No	No	Ghardaia	SD	G3	T4 N2a M0	CCI	40	15	2,67	X	X	X		Oui
67 ans	Oui	6	No	No	Oui	Ghardaia	SD	G2	T4 N1 M0	CCI	50	25	2	X	X	X	X	Oui
35 ans	No	0	Oui	No	No	Ghardaia	SG	G3	T1c N1 M0	CCI	40	16	2,5	X	X	X		Oui
42 ans	No	0	No	No	No	Souk Ahraas	SD	G2	T3 N0 M0	CCI	40	16	2,5	X	X	X		Oui
35 ans	Oui	3	No	No	No	Djelfa	SG	G2	T3 N1 M0	CCI	40	16	2,5	X	X	X		Oui
46 ans	Oui	4	No	No	No	Ghardaia	SG	G3	T2 N2 M0	CCI	50	25	2	X	X	X		Oui
58 ans	Oui	5	No	No	No	Ghardaia	SG	G3	T2 N0 M0	CCI	40	25	2	X	X	X		Oui
41 ans	Oui	2	No	No	No	Ghardaia	SG	G2	T2 N0 M0	CCI	40	16	2,5	X	X	X	X	Oui
61 ans	Oui	4	No	No	Oui	Mesila	SD	G2	T1 N1 M0	CLIS	50	25	2	X	X	X		Oui
52 ans	Oui	4	No	No	No	Ghardaia	SD	G2	T2 N0 M0	CCI	40	16	2,5	X	X	X		Oui
34 ans	Oui	3	No	No	No	Taref	SD	G3	T4a N2 M0	CCI	20	13	2	X	X	X		Oui
39 ans	No	0	Oui	No	No	Ain Defla	SG	G2	T2 N0 M0	CCI	50	23	2	X	X	X		Oui
45 ans	Oui	2	No	No	No	Ouargla	SD	G1	T4c N1 M0	CCI	40	16	2,5	X	X	X	X	Oui
45 ans	Oui	3	No	No	No	Adrar	SG	G1	T2 N0 M0	CCI	40	16	2,5	X	X	X		No
50 ans	Oui	5	No	No	No	Laghouat	SG	G2	T4d N2 M0	CCI	50	25	2	X	X	X		Oui
42 ans	Oui	2	Oui	No	No	Ghardaia	SG	G3	T3N0M0	CCI	50	23	2	X	X	X		Oui
							SD	G3	T0N2M0	CCI	20	10	2	X	X	X		Oui
43 ans	No	0	No	No	No	Ghardaia	SD	G2	T2c N0 M0	CCI	50	23	2	X	X	X	X	Oui
											50	25	2	X	X	X		Oui
57 ans	Oui	4	No	No	No	Tipaza	SD	G2	T2 N1 M0	CCI	40	16	2,5	X	X	X		Oui
62ans	Oui	5	No	No	Oui	Ghardaia	SG	G2	T2 N0 M0	CCI	40	15	2	X	X	X		Oui
49 ans	Oui	3	No	No	No	Aflou	SD	G2	T1c N1e M0	CCI	30	16	2	X	X	X		Oui
48 ans	Oui	2	No	No	No	Chelef	SD	G2	T2c N1 M0	CCI	40	16	2,5	X	X	X		Oui
58 ans	Oui	4	No	No	No	Ouargla	SG	G3	T1c N0 M3	CCI	60	16	3	X	X	X		Oui
39 ans	No	0	No	No	Oui	Ghardaia	SD	G2	T2 N1a M0	CCI	40	16	2,5	X	X	X		Oui
54 ans	Oui	4	No	No	No	Ghardaia	SG	G1	T2 N1a M0	CCI	40	16	2,5	X	X	X	X	Oui
58 ans	Oui	5	Oui	No	No	Ghardaia	SG	G3	T4N0M0	CCI	50	25	2	X	X	X		Oui
42 ans	Oui	0	No	No	No	Mostaghanem	SG	G2	T3 N0 M0	CCI	50	25	2	X	X	X		Oui
62 ans	Oui	5	No	No	No	Tiaret	SG	G2	T4 N1 M0	CCI	50	25	2	X	X	X		Oui
29 ans	No	0	No	No	No	Tiaret	SD	G2	T2N1M0	CCI	40	16	2,5	X	X	X		Oui
54 ans	Oui	4	No	No	No	Tiaret	SD	G2	T2 N0 M0	CCI	50	25	2	X	X	X	X	Oui

Figure III.1 : Exemples de données extraites de fichiers et de rapports.

III.2.1.1. Définition des variables utilisées:

Dans cette étude, on a utilisé les variables suivantes :

- **Age :** l'âge d'un patient.
- **Marie/célibataire :** Situation sociale.
- **Nombre des enfants** : Nombre de naissances.
- **La puberté précoce :** première règles avant l'âge de 12 ans.
- **Ménopause tardive :** après l'âge de 55 ans.
- **SD/ SG :** La localisation de la tumeur (SD : sein droit, SG : sein gauche)**.**
- **Grade :** Cette phase décrit la prévalence tumorale et les métastases.

- **Classification :**

  T (tumeurs):correspond à la taille de la tumeur principale et à son degré d'extension.

  N (node (ganglion)): représente le degré d'atteinte des ganglions lymphatiques situé à proximité de la tumeur.

  M (métastase): représente l'extension du cancer du sein à d'autres parties du corps, et donc, la présence éventuelle de métastases.

- **Cellule** : Cette fonctionnalité décrit le type de cellule pour la tumeur.
- **Dose total(Gy) :** représentant la dose totale du traitement.
- **Nombre de Séance :** représentant la durée du traitement.
- **Résection :** Cela signifie une mastectomie totale ou partielle.

III.2.1.2.Analyse statistique :

Les données ont été enregistrées sur une base de données Excel .Les résultats ont été exprimés en pourcentage (effectif) pour les variables qualitatives et en moyenne pour les variables quantitatives. Ils sont rapportés dans des tableaux (III-1 à III-7).

III.2.1.2.1.Caractéristique démographiques des patients :

Répartition géographique : Les patientes viennent de différentes wilayas du sud algérien avec prédominance de Ghardaïa (28%) (Voir le tableau III.1 et figure III.2).

Tableau III.1: Répartition géographique des patientes.

Villes	Nombre	Pourcentage
Autre	70	16%
Alger	7	2%
Adrar	10	2%
Tamanrasset	13	3%
El Oued	14	3%
Tiaret	15	3%
Djelfa	33	7%
Laghouat	47	10%
Ouargla	118	26%
Ghardaïa	130	28%

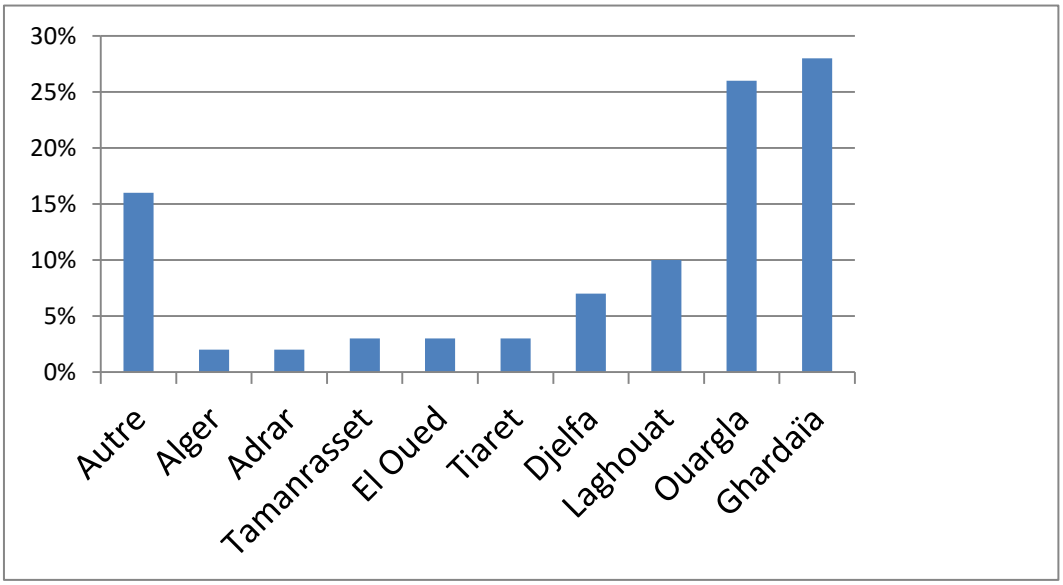


Figure III.2 : colonnes du graphique de répartition géographique.

Age : à partir le tableau III.2, on trouve que l'âge varie de 23 à 81 ans.

Tableau III.2: Répartition des patientes selon les tranches d'âge.

Age	Nombre	Pourcentage
Age<35ans	29	6%
Age entre 35-50 ans	221	48%
Age >50 ans	207	45%

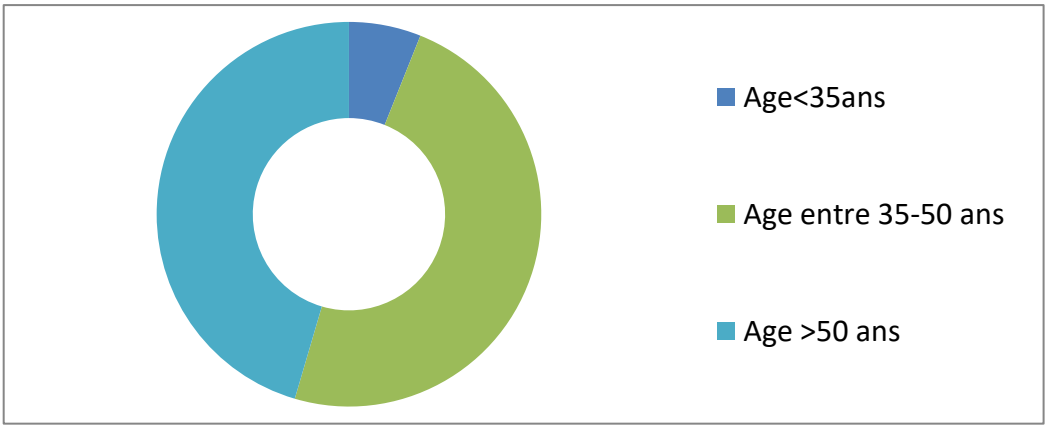


Figure III.3: cercle relatif pour l'âge d'un patient

III.2.1.2.2. La tumeur primitive :

A. <u>Aspect clinique :</u>

Localisation de la tumeur : La localisation de la tumeur primitive est répartie équitablement entre les deux seins avec d'atteinte bilatérale. Selon les données dans Tableau III.3 et figure III.4, on a trouvé qui le cancer du sein droit était plus fréquent (53%).

Tableau III.3: Répartition des patientes selon le coté du sein atteint.

Sein atteint	Nombre	Pourcentage
les deux	6	1%
Droit	243	53%
Gauche	208	46%

B. Aspect anatomopathologique :

Type histologique de la tumeur : Le type histologique le plus retrouvée est le carcinome canalaire infiltrant (CCI) dans 93 % (Tableau III.4 et la figure III.4).

Tableau III.4: Répartition des patientes selon le type histologique.

Type histologique	Nombre	Pourcentage
CCI	431	93%
Autre ((CCI+CCIS), (CCI+CCIS+Paget) (CCI+CLI),(CCIS),(CLI),(CLI+Paget),(CLIS))	32	7%

La taille tumorale : A travers le Tableau III.5 et la figure III.4, la plupart des patientes (57%) avaient une tumeur T2.

Tableau III.5: Répartition des patientes selon la taille tumorale de cancer du sein.

Taille tumorale	Nombre	Pourcentage
T0	2	1%
T1	64	14%
T2	265	57%
T3	84	18%
T4	48	10%

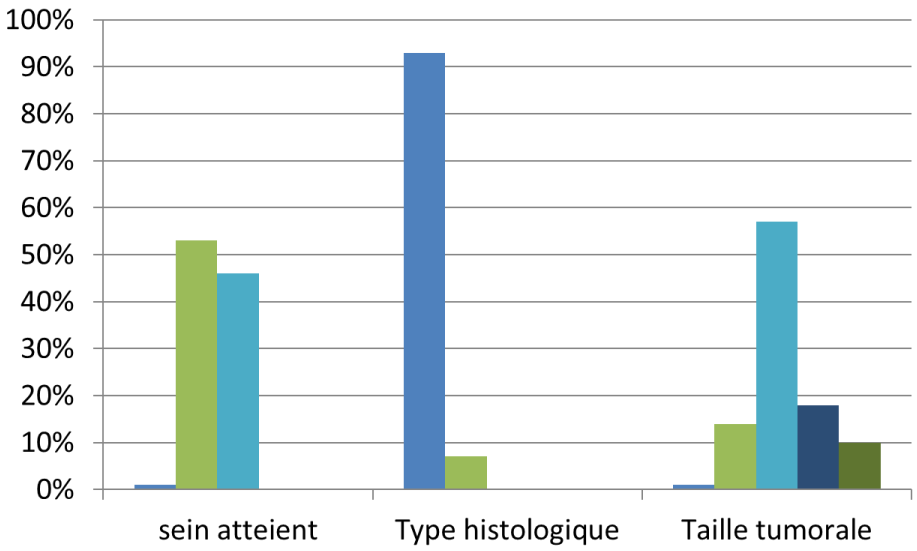


Figure III.4: colonnes du graphique des caractéristiques des tumeurs.

III.2.1.2.3.Modalité thérapeutique :

Chirurgie : La chirurgie mammaire radicale avec curage ganglionnaire axillaire était l'opération la plus pratiquée. Elle a été réalisée chez 95% (Tableau III.6 et la figure III.5).

La chirurgie mammaire conservatrice avec curage ganglionnaire axillaire était pratiquée chez 5%.

Tableau III.6: Répartition selon le type de la chirurgie.

Type de la chirurgie	Nombre	Pourcentage
Oui	442	95%
Non	21	5%

Hormonothérapie : Nombre patientes ont bénéficié de l'hormonothérapie, dans 85% des cas (Tableau III.7).

Tableau III.7: Répartition des patientes selon la prise de l'hormonothérapie.

Hormonothérapie	Nombre	Pourcentage
Oui	70	15%
Non	387	85%

Concernant la radiothérapie et la chimiothérapie : tous les patients en ont bénéficié.

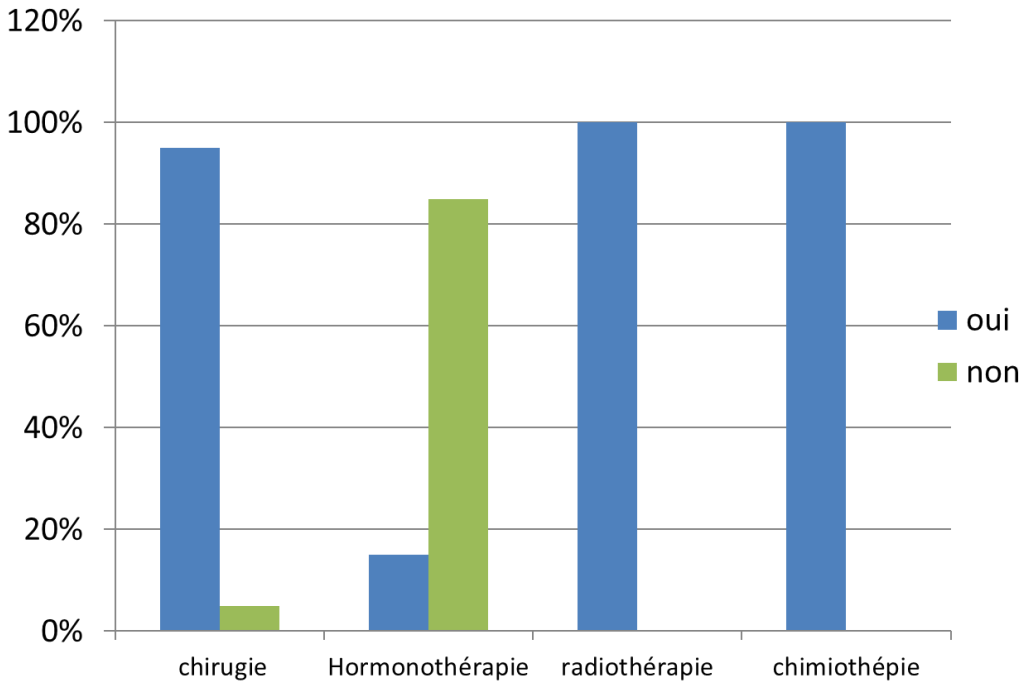


Figure III.5: colonnes du graphique des traitements.

III.3.Métriques de classification :

II.3.1.La matrice de confusion :

Dans la matrice de confusion on croise les classes cibles réelles avec les classes prédites obtenues. Ceci nous donne le nombre d'instances correctement classées et mal classées.

Tableau III.8: Matrice de confusion pour une classification binaire.

		Classes actuels	
		Positive	**Négative**
Classes prédites	**Positive**	VP	FP
	Négative	FN	VN

- **VP :** vrais positifs est le nombre d'instances positives correctement classifiées.
- **FP :** faux positifs est le nombre d'instances négatives et qui sont prédites comme positives.
- **FN :** faux négatifs est le nombre d'instances positives classifiées comme négatives.
- **VN** : vrais négatifs est le nombre d'instances négatives correctement classifiées.

À partir de la matrice de confusion on peut calculer plusieurs métriques qu'on va expliquer dans les sections suivantes [1].

III.3.2.Les métriques :

L'ensemble des métriques sont utilisées pour évaluer les méthodes d'apprentissage automatique. À partir de la matrice de confusion, de nombreuses mesures de performance du modèle peuvent être dérivées.

Accuracy : correspond à tous les modèles correctement classés divisés par le nombre total de modèles [1].

$$\textbf{Accuracy} = \frac{VP+VN}{VP+FP+FN+VN} \quad \textbf{(III.1)}$$

Précision : Cela définit l'exactitude du modèle en termes de prédiction

$$\textbf{Précision} = \frac{VP}{VP+FP} \quad \textbf{(III.2)}$$

Recall (Sensitivité): Cette mesure de performance implique comment différentes valeurs et variables indépendantes affectent une variable dépendante [2].

$$\textbf{Recall} = \frac{VP}{VP+FN} \quad \textbf{(III.3)}$$

Le score F1 (F1 score) : Peut être interprété comme une moyenne pondérée de la précision et la sensitivité, où un score F1 atteint sa meilleure valeur à 1 et son pire score à 0. Par conséquent, ce score prend en compte à la fois les cas faux positifs et les cas faux négatifs. Intuitivement, ce n'est pas aussi facile à comprendre que le taux de succès, mais F1 est généralement plus utile que le taux de succès, surtout si nous avons une distribution de classe inégale. Le taux de succès fonctionne mieux si les cas faux positifs et les cas faux négatifs ont une valeur similaire. Si la valeur des cas faux positifs et des cas faux négatifs est très différente, il est préférable d'examiner à la fois la précision et la sensitivité. Le score F1 est une métrique unique qui combine la sensitivité et la précisions en utilisant la moyenne harmonique [5].

$$\textbf{F1 Score} = \frac{2 \times (Recall \times Precision)}{(Recall + Precision)} \quad \textbf{(III.4)}$$

III.4.Métriques de régression :

Pour évaluer les modèles de régression et de les comparer, on peut calculer la distance entre les valeurs prédites et les valeurs vraies. Cela nous donne plusieurs critères :

III.4.1.L'erreur quadratique moyenne :

L'erreur quadratique moyenne RMSE (Root mean squared error) est une formule populaire pour mesurer le taux d'erreur d'un modèle de régression. Cependant, il ne peut être comparé qu'entre des modèles dont les erreurs sont mesurées dans les mêmes unités [1]. Plus la valeur de RMSE est faible, meilleure est la performance du modèle.

$$\mathbf{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n}} \qquad \textbf{(III.5)}$$

a : Cible réelle.

p : Cible prévue.

III.4.2.L'erreur absolue moyenne :

L'erreur absolue moyenne MAE (Mean Absolut error) a la même unité que les données d'origine et ne peut être comparé qu'entre des modèles dont les erreurs sont mesurées dans les mêmes unités. Son ampleur est généralement similaire à celle du RMSE, mais légèrement plus petite. a et p sont défini dans l'erreur quadratique moyenne. Une valeur de MAE plus faible est préférable.

$$\mathbf{MAE} = \frac{\sum_{i=1}^{n} | \ p_i - a_i |}{n} \qquad \textbf{(III.6)}$$

III.4.3.Le coefficient de détermination R^2 :

Le coefficient de détermination R^2 (R squared error), résume le pouvoir explicatif du modèle de régression. R^2 est calculé à partir des termes des sommes des carrés :

$$\mathbf{R^2} = 1 - \frac{SSE}{SST} \qquad \textbf{(III.7)}$$

SSE : erreur sur la somme des carrés

SST : somme des carrés total

R^2 décrit la proportion de variance de la variable dépendante expliquée par le modèle de régression. Si le modèle de régression est « parfait », SSE vaut zéro et R^2 vaut 1, Si c'est un échec total, SSE vaut SST, aucune variance n'est expliquée par la régression et R^2 vaut zéro [3]. Une valeur de R^2 plus proche de 1 indique un meilleur ajustement du modèle aux données.

III.4.4.MSE (Mean Squared Error):

Mesure le degré d'erreur dans les modèles statistiques. Il évalue la différence quadratique moyenne entre les valeurs observées et prédites. Une valeur de MSE plus faible indique une meilleure performance du modèle.

$$\mathbf{MSE} = \frac{1}{N}\sum_{N=1}^{i}(y_i - y'_i)^2 \qquad \textbf{(III.8)}$$

y_i : est la valeur observée.

y'_i : est la valeur prédite correspondante.

N : le nombre d'observations.

III.5.Langage de programmation :

Aujourd'hui, il y a différents langages de programmation et chaque langage présent ses propres particularités. Parmi ces langages, nous avons opté pour Python. Il est inventé par « Guido van Rossum », la première version de python est sortie en 1991[4]. C'est est un langage de programmation interprété, multiparadigme et multiplateformes.

Python possède de nombreux avantage :

1. **Simple d'utilisation :** Python est un langage de programmation extrêmement facile à comprendre et à assimiler, ce qui le rend accessible aux novices.
2. **Un large éventail :** Python possède une large gamme de modules et de frameworks spécialement conçus pour l'apprentissage automatique, comme TensorFlow, Scikit-learn, PyTorch, et bien d'autres encore.
3. **Flexibilité :** Python est un langage polyvalent qui facilite la collaboration entre divers outils et bibliothèques afin de concevoir des solutions personnalisées.
4. **Performances :** traiter de grandes quantités de données et exécuter de manière efficace des modèles d'apprentissage automatique.

Google colaboratory : est une plateforme offerte gratuitement par google permettant d'écrire et exécuter du code paython dans votre navigateur. Elle vous permet en particulier d'exécuter des notebooks jupyter sans avoir besoin de vous soucier de votre matériel ou des logiciels installé sur votre ordinateur. Google colab est un outil qui facilité également l'accès a des ressources de calcul et aux bibliothèques d'apprentissage automatique usuelle.

III.6.Application :

Notre modèle a été écrit à l'aide de Google Colab avec un environnement de programmation python3 en utilisant l'apprentissage automatique.

- Dans une première partie, nous avons utilisé la classification binaire pour déterminer si le patient suivra ou non un traitement hormonal.
- Dans la deuxième partie, nous avons utilisé la régression pour déterminer la dose totale (Gy).

III.6.1.Partie 1 : La classification binaire :

Dans cette partie, des modèles de classification binaire ont été essayés pour déterminer notre objectif est d'entraîner notre modèle pour prédire si un patient suivra ou non un traitement hormonal. Le modèle a été construit à l'aide des algorithmes suivants:

- Logistic Regression.
- K-Nearest Neighbors.
- Random Forest.
- Decision Tree.
- Gaussian Naïve Bayes.

III.6.1.1.La normalisation des données :

La normalisation des données est une méthode de prétraitement des données qui permet de réduire la complexité de la base de données. Elle consiste à convertir les valeurs de la base de données en nombres (voir la Figure III.6).

SD:1/ SG:0	Grade G1:1 G2:2 G3:3 G4:4	T T0:0 T1:1 T2:2 T3:3 T4:4	N N0:0 N1:1 N2:2 N3:3 N4:4	M M0:0 M1:1	Cellule CCI:0 CLI:1 CLIS:2
1	3	4	2	0	0
1	2	4	1	0	0
0	3	1	1	0	0
1	2	3	0	0	0
0	2	2	1	0	0

Figure III.6: Échantillon de données après simplification.

III.6.1.2.Etapes de classification :

Le processus de création d'un modèle d'apprentissage doit suivre les étapes suivantes en commençant par le chargement des données jusqu'à la validation du modèle.

1. Importer les bibliothèques : Nous avons utilisé les bibliothèques illustrées dans le Tableau III.9 et la figure III.7.

```
[4]: import numpy as np    # Import Numpy for data statistical analysis
     import matplotlib.pyplot as plt     # Import matplotlib for data visualisation
     import pandas as pd           # Import Pandas for data manipulation using␣
      ↪dataframes
     import seaborn as sns  # Statistical data visualization
     from sklearn.preprocessing import StandardScaler
```

Figure III.7: Importation des bibliothèques.

Tableau III.9: Les bibliothèques les plus importantes utilisées dans notre étude.

Bibliothèque	La description
Numpy	Une bibliothèque spécialisée en calcul scientifique en langage Python qui aide à travailler avec des tableaux et des données.
Pandas	aide au nettoyage et à l'organisation des données, et aussi permet de charger, de fusionner ou encore de manipuler des données.
SKlearn	Bibliothèque principale utilisée dans les projets d'apprentissage automatique. Elle contient de nombreux algorithmes et méthodes utilisés dans le domaine de l'apprentissage automatique, comme la classification, en plus de leur utilisation dans la phase de traitement des données et d'évaluation des modèles.
Seaborn	Une bibliothèque spécialisée pour les graphiques et la création d'interfaces avancées.
Matplotlib	Pour nous aider à visualiser lors de notre analyse exploratoire de l'ensemble de donnée.

2. Télécharger les Dataset :

Cet ensemble de données comprend des informations sur les patients.

```
[6]: data=pd.read_csv('/content/CancerDeSein.csv', delimiter=';')
```

```
[ ]: data.head()
```

```
[ ]:
```

	A	M	n	L	Me	De	SD	Gr	T	N	M	Ce	Se	Ds	Ho	Re	Do
0	43	1	4	0	0	0	1	3	4	2	0	0	15	2.67	0	1	40.0
1	67	1	6	0	0	1	1	2	4	1	0	0	25	2.00	1	1	50.0
2	35	0	0	1	0	0	0	3	1	1	0	0	16	2.50	0	1	40.0
3	42	0	0	0	0	0	1	2	3	0	0	0	16	2.50	0	1	40.0
4	35	1	3	0	0	0	0	2	2	1	0	0	16	2.50	0	1	40.0

```
[ ]: data.tail()
```

```
[ ]:
```

	A	M	n	L	Me	De	SD	Gr	T	N	M	Ce	Se	Ds	Ho	Re	Do
437	62	1	4	1	0	0	0	3	3	1	0	0	56	2.00	0	1	112.00
438	38	1	2	0	0	0	0	2	2	0	0	0	48	2.00	0	1	96.00
439	48	1	9	1	0	0	0	3	3	0	0	0	15	2.67	1	1	40.05
440	60	1	5	1	1	1	1	3	3	1	0	0	48	2.00	0	1	96.00
441	47	1	4	0	0	0	0	3	1	2	0	0	48	2.00	0	1	96.00

Nous avons maintenant 441 lignes et 17 colonnes.

```
In [ ]: data.shape
```

```
Out[ ]: (442, 17)
```

3. Résumé statistique :

Comprend les valeurs de décompte, moyennes, minimales et maximales ainsi que certains pourcentages.

```
Out[ ]:
```

	A	M	n	L	Me	De	SD	
count	442.000000	442.000000	442.000000	442.000000	442.000000	442.000000	442.000000	442
mean	49.187783	0.900452	3.382353	0.070136	0.061086	0.131222	0.538462	2
std	10.836804	0.299735	1.995677	0.255665	0.239759	0.338025	0.499083	0
min	23.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1
25%	41.000000	1.000000	2.000000	0.000000	0.000000	0.000000	0.000000	2
50%	48.000000	1.000000	3.000000	0.000000	0.000000	0.000000	1.000000	2
75%	56.000000	1.000000	5.000000	0.000000	0.000000	0.000000	1.000000	3
max	81.000000	1.000000	10.000000	1.000000	1.000000	1.000000	1.000000	4

Ensuite, la relation entre les moyennes des variables a été redessinée après suppression des variables qui leur sont associées. Cela a été fait à l'aide d'une carte thermique, comme le montre la figure III.8 qui affiche les résultats de la corrélation entre toutes les variables. Un

coefficient de corrélation compris entre −1 et 1 indique aucune corrélation, tandis qu'un nombre positif ou négatif indique une forte relation entre les deux variables. Pour cette raison, les instructions suivantes ont été utilisées :

```
In [ ]: plt.figure(figsize=(20,10))
        sns.heatmap(data.corr(),annot=True,cmap="plasma")
        plt.show()
```

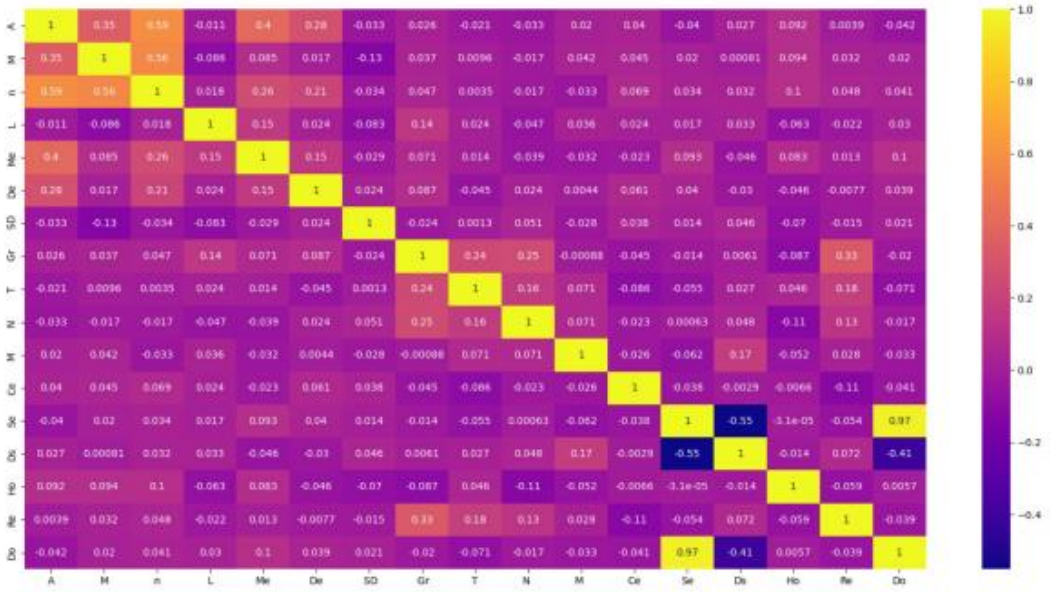


Figure III.8: Heatmap pour les variables étudiées.

4. Division des données:

Les données ont été divisées en un ensemble d'entraînement à 80 % et un ensemble de test à 20 %.

StandardScaler: Standardiser les fonctionnalités en supprimant la moyenne et en mettant à l'échelle la variance unitaire.

Enfin, les performances des algorithmes ont été évaluées à l'aide de la matrice d'incertitude.

III.6.1.3.Construire le modèle

a. Logistique régression : Implémentation d'algorithme logistique régression**:**

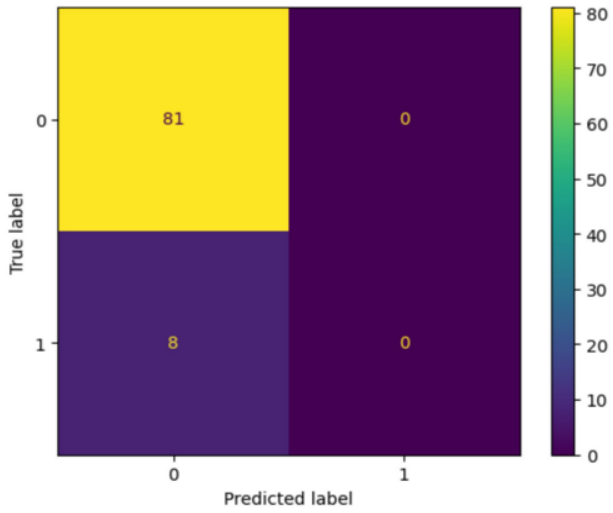


Figure III.9: Matrice de confusion d'algorithme logistique régression.

Résultats d'entraînement et de prédiction selon l'algorithme logistique régression sont donnes dans la figure III.10.

	precision	recall	f1-score	support
0	0.91	1.00	0.95	81
1	0.00	0.00	0.00	8
accuracy			0.91	89
macro avg	0.46	0.50	0.48	89
weighted avg	0.83	0.91	0.87	89

Figure III.10: Rapport de classification pour l'algorithme RL.

b. KNeighbors Clasifier : Implémentation d'algorithme KNN:

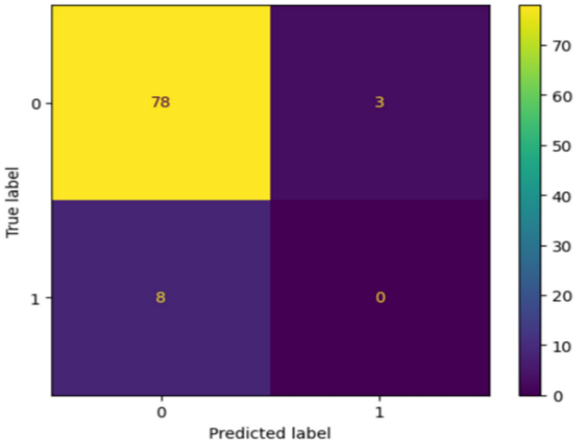


Figure III.11: Matrice de confusion d'algorithme KNN.

Résultats d'entraînement et de prédiction selon l'algorithme KNN sont indiqués dans la figure III.12.

	precision	recall	f1-score	support
0	0.91	0.96	0.93	81
1	0.00	0.00	0.00	8
accuracy			0.88	89
macro avg	0.45	0.48	0.47	89
weighted avg	0.83	0.88	0.85	89

Figure III.12: Rapport de classification pour l'algorithme KNN.

c. Random forest classifier : Implémentation d'algorithme les forêts aléatoires ou les forêts de décision aléatoires**.**

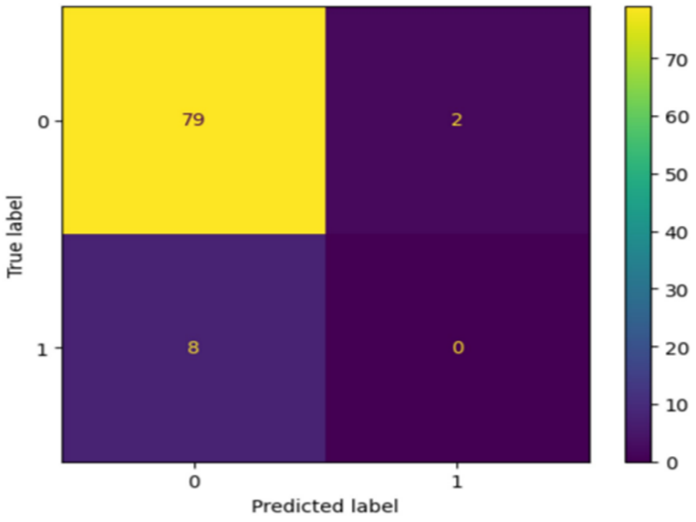


Figure III.13: Matrice de confusion d'algorithme Random forest.

Résultats d'entraînement et de prédiction selon l'algorithme Random forest sont indiqués dans la figure III.14.

	precision	recall	f1-score	support
0	0.91	0.98	0.94	81
1	0.00	0.00	0.00	8
accuracy			0.89	89
macro avg	0.45	0.49	0.47	89
weighted avg	0.83	0.89	0.86	89

Figure III.14: Rapport de classification pour l'algorithme RF.

d. Decision tree classifier : Implémentation d'Algorithme d'arbre de décision (REPTree) :

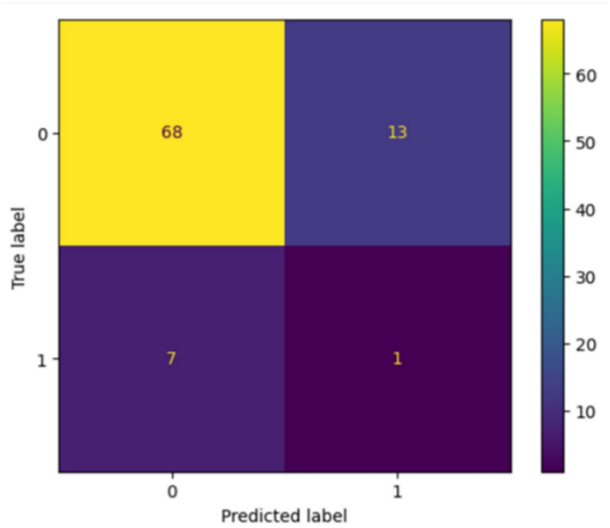


Figure III.15: Matrice de confusion d'algorithme Decision tree.

Résultats d'entraînement et de prédiction selon l'algorithme Decision tree sont indiqués dans la figure III.16.

	precision	recall	f1-score	support
0	0.91	0.84	0.87	81
1	0.07	0.12	0.09	8
accuracy			0.78	89
macro avg	0.49	0.48	0.48	89
weighted avg	0.83	0.78	0.80	89

Figure III.16: Rapport de classification pour l'algorithme DT.

e. Gaussian Naïve Bayes: Application de l'algorithme Gaussian Naïve Bayes:

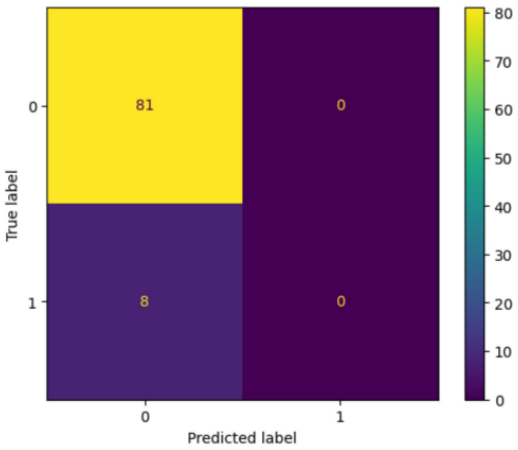


Figure III.17: Matrice de confusion d'algorithme gaussian NB**.**

La figure suivant présentent les résultats d'application de" Naïve Bayes" sur la base des données :

	precision	recall	f1-score	support
0	0.91	1.00	0.95	81
1	0.00	0.00	0.00	8
accuracy			0.91	89
macro avg	0.46	0.50	0.48	89
weighted avg	0.83	0.91	0.87	89

Figure III.18: Rapport de classification pour l'algorithme Gaussian Naïve Bayes.

III.6.1.4. Comparaison des algorithmes :

Comparaison des performances des différents algorithmes au terme de l'expérimentation de classification de données de cinq modèles différents (Naive Bayes, Logistic Regression, Decision Tree, Random Forest, Gaussian Naïve Bayes) utilisant l'apprentissage automatique.

On note que les modèles de régression logistique et de Gaussian Naïve Bayes donnent une meilleure accuracy (91%) par rapport aux trois autres algorithmes. En revanche, nous notons que le modèle Decision Tree a obtenu de mauvais résultats parmi les autres modèles.

III.6.2.Partie 2 : Regression :

Dans cette section, nous avons utilisé la prédiction pour déterminer la dose totale. Trois algorithmes ont été utilisés pour atteindre cet objectif, notamment :

- Linear regression.
- Support Vector Regression (Kernel= 'linear').
- Support Vector Regression (Kernel= 'sigmoid').
- Support Vector Regression (Kernel= 'Polynomial')
- Support Vector Regression (Kernel= 'Radial Basis Function')

Nous avons divisé nos données en deux ensembles, 80% de données pour faire l'apprentissage, et 20% de données pour le test.

En général, le choix du meilleur modèle dépend de la nature des données et de la relation entre les variables. Si les données ont des relations linéaires simples, un modèle linéaire peut être approprié. Si les données sont plus complexes et ont des relations non

linéaires, un modèle linéaire du noyau ou un modèle sigmoïde du noyau peut être plus approprié.

Pour déterminer quel modèle est le meilleur pour prédire les valeurs cibles, nous les comparons en termes de mesures d'évaluation telles que R^2, MSE, RMSE et MAE.

III.6.2.1.Appliquer l'algorithme : On peut appliquer les algorithmes en écrivant ce code :

a. Linear regression : Nous résumons les résultats dans le tableau suivant :

Tableau III.10: Résultats de l'application de modèle Linear Regression.

Modale	R^2	MSE%	RMSE%	MAE
Linear regression	0.9809	904.41	300.73	1.92

b. SVR (kernel= 'linear') : Nous résumons les résultats dans le tableau suivant :

Tableau III.11: Résultats de l'application de modèle SVR (kernel = 'linear')

Modale	R^2	MSE%	RMSE%	MAE
SVR (kernel = 'linear')	0.985	672.13	259.26	0.84

c. SVR (kernel= 'sigmoid') : Nous résumons les résultats dans le tableau suivant :

Tableau III.12: Résultats de l'application de modèle SVR (kernel = 'sigmoide').

Modale	R^2	MSE%	RMSE%	MAE
SVR (kernel='sigmoide')	-0.0288	46103.09	2147.16	13.34

d. SVR (Kernel= 'Polynomial'): les résultats obtenus sont dans le tableau suivant:

Tableau III.13: Résultats de l'application de modèle SVR (Kernel= 'Polynomial').

Modale	R^2	MSE%	RMSE%	MAE
SVR (Kernel= 'Polynomial')	0.933	2966.55	544.66	3.38

e. **SVR (Kernel= 'Radial Basis Function') :**

Les résultats obtenus sont dans le tableau suivant :

Tableau III. 14: Résultats de l'application de modèle SVR (Kernel= 'Radial Basis Function').

Modale	R^2	MSE%	RMSE%	MAE
SVR (Kernel= 'Radial Basis Function')	0.752	11110.51	1054.06	5.78

III.6.2.1.Discussion:

En regardant les résultats représenté dans les tableaux, nous constatons que le modèle le plus proche des valeurs réelles est le modèle SVR (kernel = 'linear'). Ce modèle semble très performant et capable de prédire les valeurs cibles avec une grande précision.

Linear regression, l'analyse les résultats de ce modèle montre que les performances sont globalement bonnes.

SVR (Kernel= 'Polynomial') et SVR (Kernel= 'Radial Basis Function') les résultats sont élevés pour les valeurs réelle.Cela signifie que les prédictions du modèles peuvent être éloignées des valeurs réelles.

Quant au modèle (kernel = 'sigmoide'), nous constatons que les résultats sont très élevés pour MSE, RMSE, MAE et négatifs pour R2, ce qui indique que les prédictions du modèle sont très éloignées des valeurs réelles.

III.7.Conclusion :

Notre étude dans ce chapitre nous a permis de décrire le profil des patientes qui présentent du cancer de sein (CS), les résultats étaient donc:

- L'âge moyen au moment du diagnostic de la tumeur initiale était entre 35-50 ans
- Le CCI était le type histologique le plus fréquent (93%).
- La plupart des patientes soit 57% avaient une tumeur T2, où les tumeurs localisées dans le sein droit à la majorité patients (53%).
- la résection totale était l'opération la plus pratiquée. Elle a été réalisée chez 95%.

- La radiothérapie et chimiothérapie a été réalisé dans 100 % des cas. quant au traitement hormonal, il était de 15 %.

Nous avons fait la description du dataset pour déterminer si le patient suivra ou non un traitement hormonal, après avoir appliqué cinq méthodes de performance. Ensuite, nous comparons les résultats de ces algorithmes entre eux. Pour avoir un résultat, les méthodes en affichant le rapport et la précision et f1 score de chaque modèle. Le résultat obtenu a montré que le modèle Logistic regression et Gaussian Naïve Bayes a donné de meilleurs résultats avec ce type de données. Avec une accuray de 91 %,

Après, nous avons utilisé la regression pour déterminer la dose totale (Gy), puis nous avons utilisé les données de test pour obtenir des données prédites, et comparé les résultats avec les données réelles à l'aide des mesures de performances. Pour la validation de notre modèle, nous avons utilisé plusieurs mesures de performance, telles que MSE, MAE, RMSE, R2. L'expérimentation a montré que l'utilisation des SVR (Support Vector Machine) (kernel = « linear ») a donné de meilleurs résultats (R^2=0.985, MSE=672.13%, RMSE=259.26%, MAE=0.84) par rapport les valeurs réelles.

Liste de références :

[1] د. علاء طعيمة، *التعلم الالي والتعلم العميق وعلم البيانات*، كلية علوم الحاسوب والتكنولوجيا المعلومات جامعة القادسية العراق2

[2] Rachid Mifdal, « Application des techniques d'apprentissage automatique pour la prédiction de la tendance des titres financiers », L'obtention De La Maitrise, Sous la direction de M.Edmond Miresco, École De Technologie Supérieure Université Du Québec, 2019, p 176.

[3] Saed Sayad, «Model Evaluation- Regression», from: https://www.saedsayad.com/model_evaluation_r.htm. [Consulté le 18/05/2024]

[4]"PYPL PopularitY of Programming Language index". Pypl.github.io. Archived from the original on 14 March 2017.

[5] Berrimi Mohamed, « Deep Learning for Detecting and Identifying Blinding Retinal Diseases Problematic», Thèse de master LMD en Informatique, sous la direction de Abdelouahab Moussaoui, Université Ferhat Abbas Sétif 1, 2019, 78p.

Conclusion générale

Conclusion générale

De nombreux chercheurs ont utilisé les techniques d'apprentissage automatique et d'intelligence artificielle pour la prédiction et la classification du cancer du sein.

Dans ce mémoire, nous avons présenté des modèles d'apprentissage automatique supervisé pour la classification et regrssion de cancer du sein sur l'ensemble des données que nous avons extraites des archives du service de radiothérapie de l'hôpital Mohamed Boudiaf Ouargla.

Pour la tâche de classification : Après avoir appliqué les algorithmes différents (Naive Bayes, Logistic Regression, Decision Tree, Random Forest, Gaussian Naïve Bayes) pour déterminer si le patient subira ou non un traitement hormonal. La régression logistique et le Bayes naïf gaussien ont donné le meilleur résultat avec une accuracy (91 %), tandis que le modèle d'arbre de décision était faible parmi les autres. Avec une accuracy (78%).

Pour la régression afin de prédire la dose nécessaire au traitement, nous avons utilisé : Linear regression, SVR (kernel ='linear'), SVR (kernel ='sigmoide'), SVR (Kernel= 'Polynomial') et SVR (Kernel= 'Radial Basis Function'). Pour les valeurs réelles, SVR (kernel ='linear') a donné de meilleurs résultats. Ce modèle semble très performant et capable de prédire les valeurs cibles avec une grande précision. Quant au modèles SVR (kernel ='sigmoide'), SVR (Kernel= 'Polynomial') et SVR (Kernel= 'Radial Basis Function') a donné des valeurs éloignées de la valeur réelle moyenne, cela signifie que les modèles ne parvient pas à interpréter les données.

Par conséquent le modèle logistic et gaussian NB peut être utilisé pourpour déterminer si le patient suivra ou non un traitement hormonal et SVR (kernel ='linear') pour déterminer la dose totale ce qui aide grandement les médecins à établir un diagnostic approprié.

Et nous souhaitons peut améliorer dans le prochain avenir.