République Algérienne Démocratique et Populaire

Ministère de l'enseignement supérieur et de la recherche scientifique

Ecole Nationale Supérieure de Statistique et d'Economie Appliquée

Département de statistique appliquée et d'économétrie

Mémoire de fin d'étude en vue de l'obtention du diplôme de Master en Statistique Appliquée

Thème :

Etude des facteurs de risque du cancer du sein selon une approche de Data Mining

Présenté par : HENIA Khaled **Encadré par :** Dr MEDJTOUH.F

Année universitaire 2020-2021

Dédicaces

Je dédie ce modeste travail :

A mes très chers parents CHETBI Zahira et Abdelmalek pour leurs amours, sacrifices et leurs soutiens.

A mon très cher frère Marouane et à ma petite adorable sœur Hadjer.

A ma grand-mère CHETIBI Saada et à la mémoire de mes grands-parents HENIA Mohamed et CHETIBI Guemera. Et à la mémoire de mon très cher grand-père CHETIBI Bachir, qui nous a quitté trop tôt, et dont le mentorat et la sagesse me manquent.

Et à tous mes amis de l'ENSSEA pour les bons moments qu'on a passés ensemble et qui resteront gravé à tout jamais.

H. Khaled

Remerciements

En premier lieu, je tiens à exprimer toute ma reconnaissance à ma directrice de mémoire, Madame MEDJTOUH. Je la remercie de m'avoir encadré, orienté, aidé et conseillé.

Je désire aussi remercier les professeurs de l'Ecole Nationale Supérieure de Statistique et d'Economie Appliquée (ENSSEA), qui m'ont fourni les outils nécessaires à la réussite de mes études universitaires.

Et j'adresse aussi mes remerciements à tous ceux qui m'ont aidé et encouragé de près ou de loin à faire ce travail.

H. Khaled

❖ Liste des abréviations :

AD : Arbre de décision.
CART : Classification And Regression Tree.
CCI : Carcinome canalaire infiltrant.
CHAID : CHi-squared Automatic Interaction Detector.
CIRC : le Centre International de Recherche sur le Cancer.
CLI : Carcinome lobulaire infiltrant.
ECD : Extraction des Connaissances à partir des Données.
GLOBOCAN : Global Cancer Observatory.
GRC : Gestion de la relation client.
IMC : Indice de masse corporelle.
KDD : Knowledge discovery in databases.
MNT : Maladies Non Transmissibles.
THM : Traitement hormonal de la ménopause.

❖ Liste des tableaux :

❖ Liste des figures :

❖ Liste des annexes :

Résumé

L'objectif de notre étude est de construire un classifieur bayésien naïf et un arbre de décision, d'un côté pour prévenir le cancer du sein à l'aide des facteurs de risque de cette maladie, qui est le cancer le plus fréquent chez la femme dans le monde et en Algérie, et d'un autre pour connaitre les facteurs de risque les plus discriminants.

A l'aide des résultats obtenus sur notre échantillon de 247 femmes, on peut conclure que, pour notre échantillon l'âge est le facteur de risque le plus discriminant et l'allaitement au sein est un facteur protecteur et le risque de développer un cancer du sein diminuera encore plus si la durée d'allaitement est plus d'un an et un Indice de la Masse Corporelle élevé augmentera aussi le risque de développer un cancer du sein.

Les résultats des modèles obtenus sont acceptables, un arbre de décision avec 84% de précision et 80% pour notre classifieur bayésien naïf.

Mots-clés : Classifieur bayésien naïf, arbre de décision, facteurs de risque, cancer du sein, cancer, Inde de la Masse Corporelle.

Abstract

The objective of our study is to build a naive Bayesian classifier and a decision tree, on one hand to prevent breast cancer with the help of the risk factors of this disease, which is the most common cancer in women in the world and in Algeria, on the other hand to find out the most discriminating risk factor for breast cancer.

Using the results obtained on our sample of 247 women, we can conclude for our sample, that the age is the most discriminating risk factor and breastfeeding is a protective factor and the risk of developing breast cancer will decrease further more if the duration is more than a year and a high Body Mass Index will increase also the risk of developing breast cancer.

The results of the models obtained are acceptable, a decision tree with 84% precision rate and 80% for our naive Bayesian classifier.

Keywords: Naive Bayesian classifier, decision tree, risk factors, breast cancer, cancer, Body Mass Index.

❖ Sommaire :

Introduction générale

Aujourd'hui la femme joue un rôle très important dans la société, elle est non seulement celle qui donne la vie, qui assure l'éducation des enfants et qui s'occupe du foyer, mais aussi celle qui est au service de son peuple, et qui attribue à la construction de son pays, c'est pourquoi nous devons prendre soin d'elle et préserver sa santé, et au moment actuelle peu de femmes qui jouissent d'un bon état général de santé et souffrent de plusieurs maladies. Et selon l'organisation mondiale de la santé les maladies non transmissibles constituent la plus grande cause de décès chez les femmes dans le monde, avec 18,9 millions de décès en 2015.[1]

Parmi ces maladies non transmissibles, on trouve le cancer du sein qui est le cancer le plus dangereux chez la femme, qui est une maladie de cellules, qui se transforment en cellules anormales et se multiplient de façon incontrôlée au niveau du sein.

Le cancer du sein est de loin le cancer le plus fréquent chez la femme dans le monde avec environ 2,2 millions de nouveaux cas diagnostiqués en 2020, et de 685 000 femmes mortes à cause de cette maladie, c'est ce qui a fait de cette maladie la première cause de mortalité chez la femme, et selon l'OMS près d'une femme sur 12 développe un cancer du sein.[2]

Le cancer du sein est un véritable problème de santé publique en Algérie, son incidence ne cesse d'augmenter avec 12 536 nouveaux cas c'est ce qui a fait de cette maladie le cancer le plus fréquent en 2020, avec 4 116 morts (2ème cancer au niveau de mortalité après le cancer du poumon, pour les deux sexes)[3], et selon des estimations faites par l'Organisation Mondiale de la Santé ces chiffres ne

[1] Organisation mondiale de la santé (OMS), disponible sur : <https://www.who.int/fr/news-room/fact-sheets/detail/women-s-health> (Consulté le 01/09/2021 à 23 :41).
[2] Selon des statistiques disponibles sur : <https://www.who.int/fr/news-room/fact-sheets/detail/breast-cancer> (Consulté le 02/09/2021 à 00 :39).
[3] Globocan 2020, disponible sur : <https://gco.iarc.fr/today/fact-sheets-populations> (Consulté le 02/09/2021 à 09 :20).

cesserons pas à augmenter et à l'horizon de 2030 l'Algérie connaîtra une augmentation de 26,6% de patientes atteintes d'un cancer du sein.[1]

Comme tout autre type de cancer, divers facteurs de risque peuvent avoir une influence sur la venue d'un cancer du sein, c'est pourquoi il est important de connaître l'influence de ces facteurs-là, et les précautions et les examens simples qui permettent de dépister un éventuel cancer du sein le plus tôt possible et donc d'accroître les chances de guérison totale.

Et dans ce contexte-là, l'objectif de notre étude est de construire des modèles de classifications qui nous permettent de classifier les femmes si elles sont atteintes d'un cancer du sein ou bien non, et de déterminer le risque de développer un cancer du sein afin de sensibiliser les femmes à passer le dépistage pour détecter si elle a un cancer du sein avant qu'il soit à un stade avancé, car la découverte de ce cancer à un stage peu avancé augmentera les possibilités de guérison, comme tout autre type de cancer. Et cela en basant sur les facteurs de risque de cette maladie, et de voir lequel de ces facteurs explique le plus cette maladie.

Pour réaliser notre étude, nous essaierons de répondre à la question principale :

Peut-on construire un modèle de classification précis qui aide à la détection du cancer du sein à partir de ses différents facteurs de risque ?

Et la recherche de la réponse de cette problématique, nous mène à poser quelques questions secondaires :

› Qu'est-ce qu'un cancer du sein et quel est le facteur de risque le plus important de cette maladie ?

[1] The Global Cancer Observatory (GCO), Cancer tommorow : Algeria, disponible sur : <https://gco.iarc.fr/tomorrow/en/dataviz/tables?group_populations=1&populations=12&mode=cancer&multiple_populations=0&years=2030> (Consulté le 02/09/2021 à 09 :22).

› Qu’est-ce qu’une classification et quelles sont ces différentes méthodes ?

› Quelle est la méthode de classification qui donne de meilleures performances ?

Afin de répondre à ces sous questions, un ensemble d’hypothèses a été posé :

› Le cancer du sein est une maladie cellulaire qui touche le sein, et l’âge et les antécédents familiaux sont les facteurs le plus importants.

› Une classification est une méthode de prédiction de classe selon un ensemble de caractéristiques, et il existe plusieurs différentes méthodes.

› L’arbre de décision donnera de meilleurs résultats que le classifieur bayésien naïf.

Pour construire ces modèles-là, et tester les hypothèses précédentes, une construction d’une base de données et la collecte de données dans le Centre Pierre-et-Marie-Curie (CPMC) est nécessaire, mais avec la situation actuelle des hôpitaux algériens durant la pandémie de COVID-19, nous n’avons pas pu faire une étude réelle sur le terrain, et pour cela une autre alternative a été trouver, une enquête en ligne a été faite pour construire une base de données afin de construire ces modèles-là et de répondre aux questions précédentes.

CHAPITRE 01 : Généralités sur le cancer du sein

Introduction

En Afrique du Nord et au Moyen-Orient, le cancer du sein est le premier cancer de la femme. Il représente 14 à 42 % de tous les cancers féminins avec une augmentation exponentielle. Son incidence standardisée selon l'âge varie entre 9,5 et 54 pour 105 femmes. Ces éléments font que l'Organisation mondiale de la santé considère le cancer du sein comme une priorité de santé publique et un problème majeur chez la femme dans cette région du monde.[1]

Pour mieux comprendre le cancer de manière générale et en particulier le cancer du sein nous avons partagé ce chapitre en trois sections :

Dans la première section « Généralités sur le cancer » nous avons défini le cancer de manière générale en présentant ses principaux types et nous avons visualiser l'épidémiologie du cancer.

Dans la deuxième section, nous avons bien défini le cancer du sein avec ses différents types et ses facteurs de risque.

Et dans la troisième et la dernière section, nous avons expliqué comment le cancer du sein est diagnostiqué et traité.

[1] Épidémiologie des cancers du sein de la femme jeune en Afrique du Nord, p 56-57.

Section 1 : Généralités sur le cancer

Le cancer est un groupe de maladies dangereuses qui tuent et qui posent un véritable problème de santé à l'échelle nationale et internationale.

1. Définition du cancer

Définition 1 : « Cancer » est un terme général appliqué à un grand groupe de maladies qui peuvent toucher n'importe quelle partie de l'organisme. L'une de ses caractéristiques est la prolifération rapide de cellules anormales qui peuvent essaimer dans d'autres organes, formant ce qu'on appelle des métastases.[1]

Définition 2 : Le cancer est une maladie génétique, c'est-à-dire qu'il est causé par des changements dans les gènes qui contrôlent le fonctionnement de nos cellules, en particulier la façon dont elles se développent et se divisent.[2]

2. Les différents types du cancer

Sur le plan médical, le mot « cancer » désigne un groupe de maladies très différentes les unes des autres. C'est pourquoi on ne devait pas parler du cancer, mais des cancers, au pluriel.
Il existe plusieurs types de cancers qui sont déterminés en fonction de l'histologie, autrement dit de la nature du tissu dans lequel ils se développent. Ainsi, on distingue :[3]

Les carcinomes

Les cellules cancéreuses apparaissent dans un épithélium, c'est-à-dire un tissu recouvrant les surfaces internes (tissu de revêtement des organes) ou les surfaces externes (épiderme par exemple). Dans cette catégorie, on distingue les

[1] Organisation mondiale de de la santé (OMS), **Thème de santé : Cancer**, disponible sur : <https://www.who.int/topics/cancer/fr/> (Consulté le 07/05/2021 à 16 :58).
[2] National cancer institute (NIH), **About cancer**, disponible sur : <https://www.cancer.gov/about-cancer/understanding/what-is-cancer> (Consulté le 07/05/2021 à 00 :48).
[3] Institut national du cancer (France), **guide patients (J'ai un cancer : comprendre et être aidé)**, p 12, disponible sur : <https://www.e-cancer.fr/Expertises-et-publications/Catalogue-des-publications/J-ai-un-cancer-comprendre-et-etre-aide> (consulté le 08/05/2021 à 15 :01).

adénocarcinomes qui se développent à partir de l'épithélium d'une glande telle que le sein et la prostate.[1]

Les sarcomes

Les cellules cancéreuses apparaissent dans un tissu de soutien ou conjonctif comme les os, la graisse ou les muscles.[2]

Les cancers hématopoïétiques ou hématologiques

Les cellules cancéreuses apparaissent dans la moelle osseuse qui fabrique les cellules du sang (globules rouges et blancs et plaquettes) et leurs précurseurs. Elles peuvent également apparaître dans les autres organes lymphoïdes (thymus, ganglions lymphatiques, rate, amygdales…). Il existe trois familles de cancers hématologiques : les leucémies, les myélomes et les lymphomes.[3]

3. Epidémiologie des cancers féminins

Que-ce que c'est une épidémiologie ?

L'épidémiologie étudie les variations de fréquence des maladies dans les groupes humains et recherche les déterminants de ces variations. Elle vise en particulier à la recherche des causes des maladies et à l'amélioration de leurs traitements et moyens de prévention.[4]

Aperçu mondiale des cancers féminins

Incidence

Durant 2020, environ 20 millions de nouveaux cas de cancéreux ont été enregistré pour les deux sexes, et parmi eux environs 9,3 millions sont des

[1] Institut national du cancer (France), comprendre prévenir et dépister, **types et stades des cancers**, disponible sur : <https://www.e-cancer.fr/Comprendre-prevenir-depister/Qu-est-ce-qu-un-cancer/Types-et-stades-des-cancers#toc-types-de-cancers> (consulté le 08/05/2021 à 17 :22).
[2] Ibid.
[3] Ibid.
[4] Alain-Jacques Valleron**, L'épidémiologie humaine Conditions de son développement en France, et rôle des mathématiques**, Académie des sciences : Rapport Science et Technologie, p 3-4.

femmes, et les cancers les plus courants chez les femmes (en termes de nombre de cas recensés) étaient les suivants (classés par type) :[1]

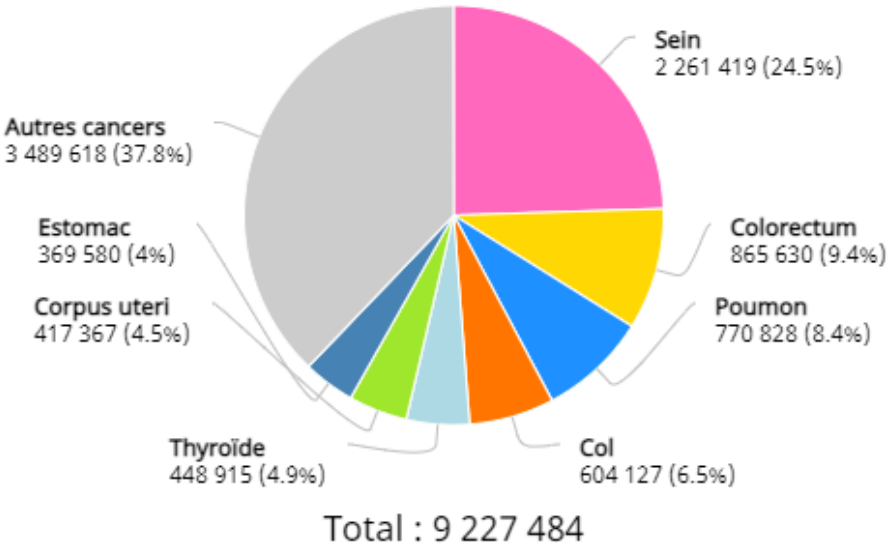

Figure (I-1) : Nombre estimé de nouveaux cas dans le monde, en 2020, pour les femmes, tous âges confondus.

Source : GLOBOCAN 2020.

Selon les estimations de l'OMS, le cancer du sein féminin est devenu le type de cancer le plus couramment diagnostiqué dans le monde, avec un total de 2,3 millions de cas ont été confirmés, dépassant pour la première fois le nombre de nouveaux cas de cancer du poumon (pour les deux sexes), et en 2020, ce cancer représentait 24,5% de tous les nouveaux cas de cancer chez les femmes dans le monde. Ont suivi le cancer colorectal (9,4%), le cancer du poumon (8,4%) et le cancer du col utérin (6,5%).[2]

Ces chiffres-là ont fait du cancer du sein un problème de santé majeur dans le monde et dans de nombreux pays où cette maladie est considérée comme le cancer le plus fréquent chez la femme dans ces pays :

[1] Organisation mondiale de la santé, centre des médias, principaux repères, détail**, cancer**, disponible sur : <https://www.who.int/fr/news-room/fact-sheets/detail/cancer> (Consulté le 18/09/2021 à 17 :02).

[2] Organisation des nations unies (ONU) : ONU info, l'actualité mondiale, disponible sur : <https://news.un.org/fr/story/2021/02/1088502> (Consulté le 18/09/2021 à 11 :06).

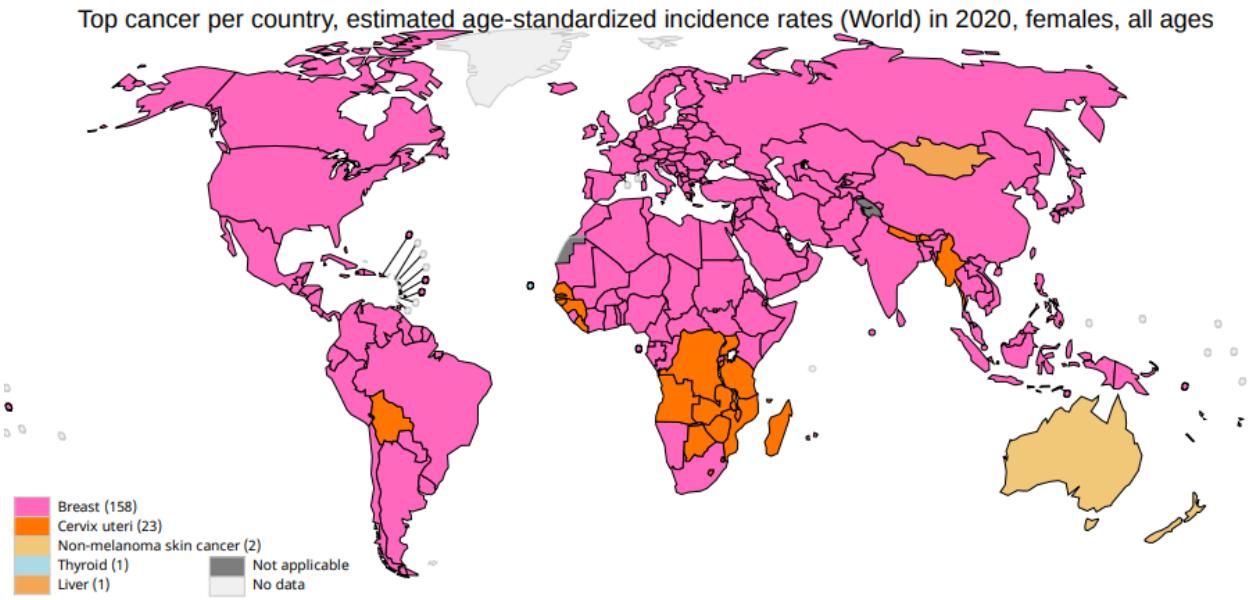

Figure (I-2) : Le cancer le plus fréquent chez la femme par pays (2020), tous âges confondus.

Source : GLOBOCAN 2020.

Mortalité

Selon le centre international de recherche sur le cancer, en 2020 le cancer est la deuxième cause de décès dans le monde, avec environ 10 millions de morts pour les deux sexes. Près d'un décès sur six est dû au cancer à l'échelle mondiale, et parmi les 10 millions de morts environ 4,5 millions sont des femmes,[1] et qui sont réparties selon le type de cancer comme suit :

[1] Organisation mondiale de la santé, centre des médias, principaux repères, détail**, cancer**, disponible sur : <https://www.who.int/fr/news-room/fact-sheets/detail/cancer> (consulté le 08/05/2021 à 19 :30).

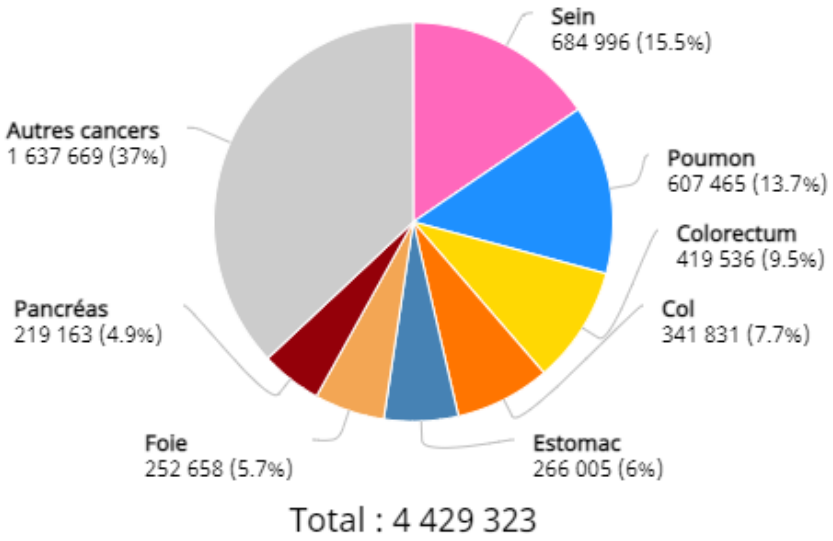

Figure (I-3) : Nombre estimé de décès en 2020 dans le monde, pour les femmes, tous âges confondus.

Source : GLOBOCAN 2020.

D'après ce graphe ci-dessus, on trouve que principalement le cancer du sein (15,5%) est le premier (parmi les cancers) responsable des décès chez la femme dans le monde.

Aperçu des cancers en Algérie

Selon le centre international de recherche sur le cancer (CIRC), pour une population de 43,8 millions, en 2020 l'Algérie a enregistré 58 418 nouveaux cas et 32 802 morts, pour les deux sexes.[1]

Presque 54% de ces nouveaux cas sont des femmes, soit 31 090 femmes, réparties selon le diagramme à barres et classées par types de cancer comme suit :

[1] Disponible sur : <https://gco.iarc.fr/>.

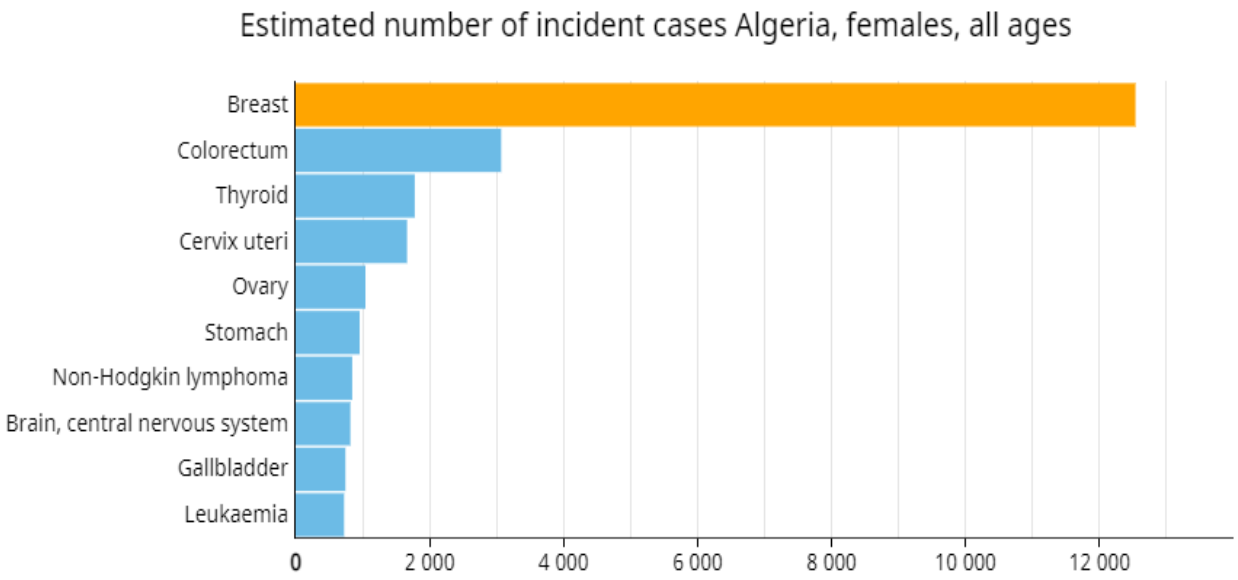

Figure (I-4) : Nombre estimé de nouveaux cas, pour les femmes en Algérie, tous âges confondus (2020).

Source : GLOBOCAN 2020.

D'après le diagramme ci-dessus, les principaux types de cancer féminin en Algérie sont les suivants :

- Le cancer du sein : 12 536 cas (40,3%).
- Le cancer colorectal : 3 068 cas (9,9%).
- Le cancer de la thyroïde : 1 778 cas (5.7%).
- Le cancer du col utérin : 1 663 cas (5.3%).
- Le cancer de l'ovaire : 1 042 cas (3.4%).
- Et 11 003 cas (35.4%) pour les autres cancers.

Pour la mortalité des cancers féminins, l'Algérie a enregistré presque 15 000 femmes mortes à cause d'un cancer, et les cancers les plus mortels chez la femme algérienne en 2020 sont les suivants :

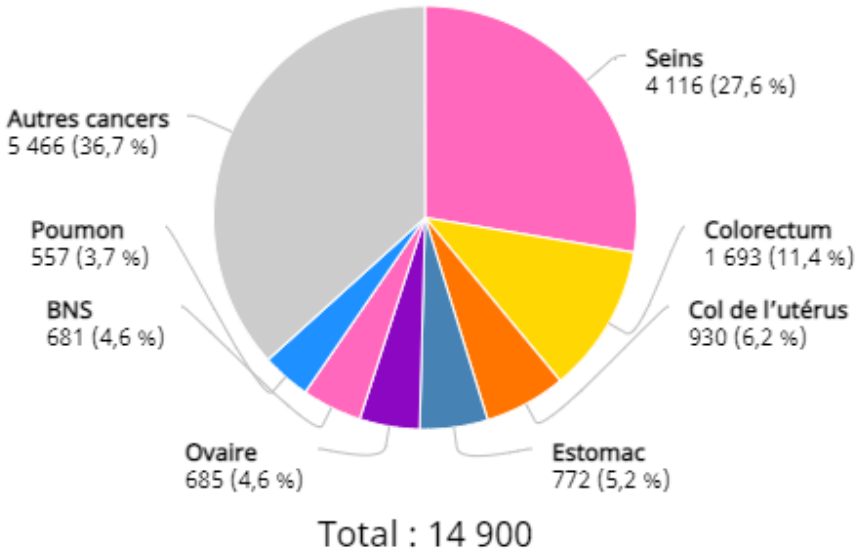

Figure (I-5) : Nombre estimé de décès par type de cancer (femme), Algérie, 2020.

Source : GLOBOCAN 2020.

Parmi les 14 900 décès (femme), 27,6% (soit 4 116 décès) sont mortes à cause d'un cancer du sein, et c'est ce qui fait du cancer du sein la première cause de mortalité (parmi les cancers) chez la femme algérienne en 2020.

Evolution des cancers en Algérie

A partir des années 90, l'Algérie a connu une transition démographique profonde et rapide qui a entraîné une modification structurelle du prol épidémiologique de sa population. Celle-ci a connu une baisse de la mortalité générale qui a été divisée par 4 en l'espace de 50 ans (16,45 pour mille à la fin des années 60 à 4,41 pour mille habitants en 2008) et une baisse importante de la mortalité infanto-juvénile corrélée à une augmentation progressive de l'espérance de vie estimée à 25 années au cours des 50 dernières années, ce qui a eu pour conséquence un vieillissement progressif de la population avec une part de plus en plus importante des personnes âgées de plus de 60 ans dans la pyramide des âges.[1]

Par ailleurs, une modification profonde du mode de vie collectif et individuel (augmentation du tabagisme, du stress, de la sédentarité, de l'urbanisation.) et d'un mode alimentaire déséquilibré sont à l'origine de l'émergence des Maladies

[1] Plan national : CANCER (2015-2019), Algérie, Octobre 2014, p 17.

Non Transmissibles (MNT) dont le cancer. Ces maladies constituent aujourd'hui plus de 80% des causes de maladies et ont en commun un certain nombre de facteurs de risque d'où la nécessité d'une politique commune de prévention contre ceux-ci.[1]

L'augmentation de l'incidence de cette maladie qui est passée de 80 nouveaux cas pour 100.000 habitants en 1990 à plus de 130 nouveaux cas pour 100.000 habitants en 2010 est significative et il est prévisible qu'elle va progresser, pouvant atteindre rapidement 50.000 cas par an (Figure I-6)[2], et ces chiffres-là ont atteint les 58 418 nouveaux cas en 2020.

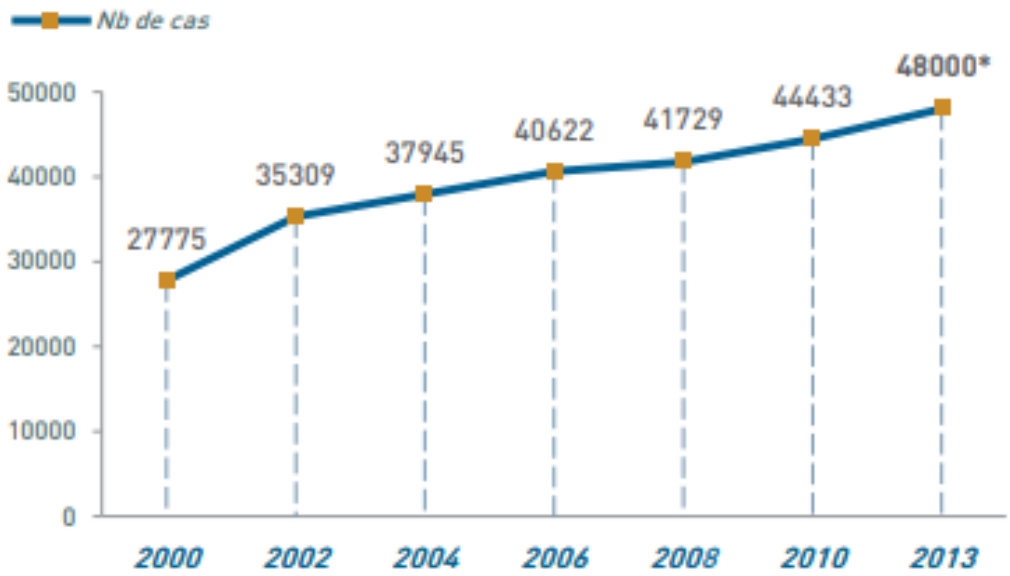

Figure (I-6) : Evolution estimée du nombre de cancer en Algérie (2000-2013).

Source : Plan national CANCER (2015-2019).

[1] Ibid.
[2] Ibid., p 18.

Evolution des cancers féminins en Algérie

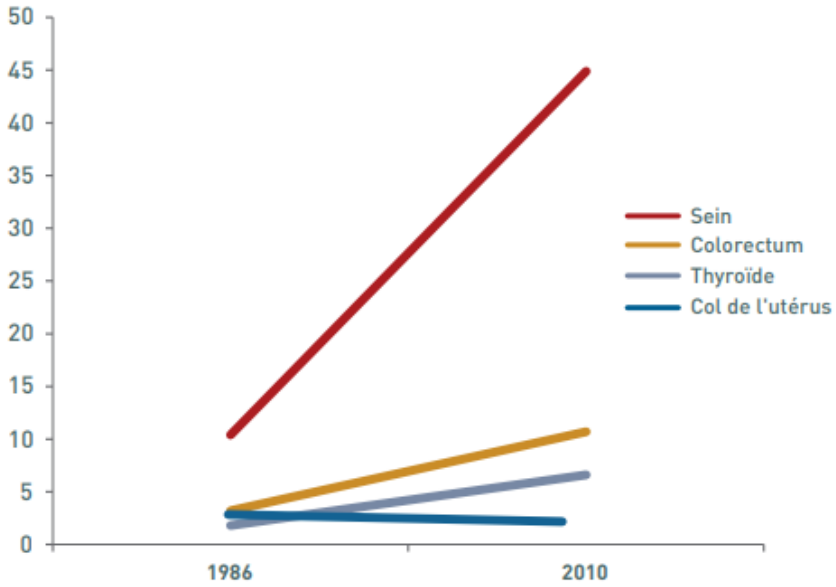

Figure (I-7) : Evolution des cancers chez les femmes entre 1986 et 2010, Algérie (nouveaux cas pour 100 000 femmes).

Source : Registre des cancers (Sétif).

D'après le graphe précédent, on remarque que l'évolution de l'incidence du cancer du sein en Algérie entre la période 1986 et 2010 est beaucoup plus importante qu'aux autres évolutions, 10 nouveaux cas pour 100 000 femmes en 1986, et en 2010, 54 nouveaux cas pour 100 000 femmes.[1]

•Nous remarquons dans les figures et les analyses précédentes que le cancer du sein est le cancer le plus fréquent chez la femme en Algérie et dans le monde, et son incidence ne cesse pas d'augmenter ce qui nous a poussé à choisir le cancer du sein comme thème de recherche.

[1] Ibid., p 21.

Section 2 : Le cancer du sein et ses facteurs

En étant le cancer le plus fréquent chez les femmes dans le monde et en l'Algérie comme nous l'avons vu précédemment, le cancer du sein est considéré comme un véritable problème majeur de santé publique.

1. Définition de cancer du sein

Définition et anatomie du sein

Les seins, du latin "sinus" qui signifie "courbure, sinuosité, pli" sont des organes pairs présents chez la femme et sous une forme atrophiée chez l'homme.[1] Le sein est constitué de graisse, de tissu conjonctif, de glandes et de canaux. Chaque sein repose sur un large muscle du thorax appelé "muscle grand pectoral". Et Il est composé de différentes parties :[2]

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.

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.

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

Le mamelon désigne la région située au centre de l'aréole et d'où sort le lait à une extrémité. Le mamelon est fait de fibres musculaires.

[1] Le journal des femmes Santé, Examens, **Sein : anatomie, examens et maladies**, Disponibles sur : <https://sante.journaldesfemmes.fr/fiches-anatomie-et-examens/2571039-sein-anatomie-examens-et-maladies/ > (Consulté le 19/05/2021 à 10 :59).
[2] Ibid.

L'aréole est la surface ronde 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.

Pour mieux visualiser cette structure, on peut imaginer un arbre avec plusieurs branches (les canaux) rattachées à un point central (le mamelon). Aux minuscules extrémités des branches se trouvent les lobules.[1]

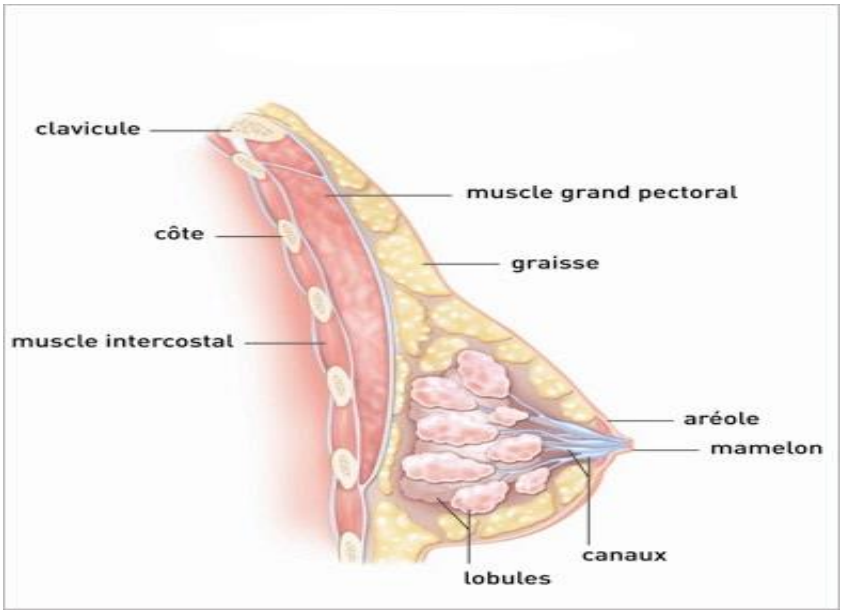

Figure (I-8) : Structure du sein.

Source : Institut national Français du cancer.[2]

Qu'est-ce qu'un cancer du sein ?

Définition 1 : Le cancer du sein est une tumeur maligne qui se forme à partir de la croissance incontrôlée de cellules mammaires anormales. Les tumeurs malignes peuvent envahir et détruire les tissus environnants et se propager à d'autres parties du corps.[3]

Définition 2 : Le cancer du sein résulte de la transformation cancéreuse d'une cellule glandulaire du sein. Cette cellule va se multiplier anarchiquement au niveau

[1] Institut national du cancer (France), **guide d'information : Comprendre le cancer du sein**, (2007), p 9.
[2] Disponible sur : <https://www.e-cancer.fr/Patients-et-proches/Les-cancers/Cancer-du-sein/Anatomie-du-sein>. (Consulté le 19/05/2021 à 17 :11).
[3] JAMES N. PARKER, M.D. ET PHILIP M. PARKER, PH. D, **Breast Cancer: A Bibliography and Dictionary for Physicians, Patients, and Genome Researchers**, (2007), p 3.

des canaux galactophores du sein (carcinome canalaire) ou des lobules du sein (carcinome lobulaire) pour ce qui est des carcinomes dits « in situ ». L'atteinte peut s'étendre aux tissus environnants, on parlera de carcinome « infiltrant ». Dans ce second cas, il est possible que les cellules cancéreuses se propagent aux ganglions lymphatiques axillaires, on parlera d'envahissement ganglionnaire.[1]

2. Types de cancer du sein

On distingue différents types de cancers du sein, selon le type de cellules à partir desquelles ils se forment. Les plus fréquents (95 %) se développent à partir des cellules des canaux (cancer canalaire) et des lobules (cancer lobulaire). Et chaque type peut être soit invasif, soit in situ.[2]

Le cancer du sein in situ (non infiltrant)

Les cellules cancéreuses ne se sont pas infiltrées dans les tissus avoisinants. Elles restent localisées au niveau des lobules du sein et des canaux galactophores. Les ganglions lymphatiques ne sont donc pas touchés, ni d'ailleurs les autres parties du corps.[3]

Le cancer du sein invasif (infiltrant)

Un cancer du sein est dit « infiltrant » si des cellules cancéreuses sont présentes dans les tissus qui entourent les lobules ou les canaux du sein où s'est formée la tumeur. Et il existe de nombreux types de cancers du sein invasifs et voici les types les plus courants :[4]

[1] FOUCAUT Aude-Marie, Thèse de doctorat : L'Activité Physique Adaptée en sénologie : des preuves scientifiques à la mise en œuvre de programmes auprès des patientes atteintes de cancer du sein, Université de Lyon, 2013, p 18-19.
[2] Europa donna France- Association contre le cancer du sein, Les différent type de cancer du sein, disponible sur : <http://www.europadonna.fr/le-cancer-du-sein/le-cancer-du-sein/differents-types-de-cancer/>. (Consulté le 04/06/2021 à 20 :20).
[3] Ooreka santé, Types de cancer du sein, disponible sur : <https://cancer-du-sein.ooreka.fr/comprendre/types-cancer-du-sein>. (Consulté le 04/06/2021 à 20 :10).
[4] Le journal des femmes Santé, disponible sur : <https://sante-medecine.journaldesfemmes.fr/faq/37060-cancer-du-sein-infiltrant-non-metastatique-traitements> (Consulté le 04/06/2021 à 20 :01).

Le carcinome canalaire infiltrant (CCI) est le type de cancer du sein infiltrant le plus observé. La tumeur initiale traverse la paroi des canaux et envahit les tissus mammaires adjacents.
Et Le carcinome lobulaire infiltrant (CLI) représente lui environ 10 % des cancers du sein infiltrants. Les cellules cancéreuses se multiplient depuis le lobule, jusqu'à coloniser les tissus mammaires avoisinants.

3. Les symptômes du cancer du sein

On appelle symptômes d'une maladie, toute manifestation anormale provoquée par cette maladie. Et parmi ces symptômes on a :[1]

Une boule dans un sein

Une boule ou une masse dans un sein est le signe d'un cancer du sein le plus couramment observé. Cette masse, en général non douloureuse, est le plus souvent de consistance dure et présente des contours irréguliers. Elle apparaît par ailleurs comme « fixée » dans le sein.

Des ganglions durs au niveau de l'aisselle (sous le bras)

Une ou plusieurs masse(s) dures à l'aisselle signifient parfois qu'un cancer du sein s'est propagé aux ganglions axillaires. Les ganglions restent toutefois indolores.

Des modifications de la peau du sein et du mamelon

- Une modification de la peau : rétraction, rougeur, œdème ou aspect de peau d'orange.

- Une modification du mamelon ou de l'aréole (zone qui entoure le mamelon) : rétraction, changement de coloration, suintement ou écoulement.

[1] Institut national du cancer (France), cancer du sein, symptômes, disponible sur : <https://www.e-cancer.fr/Patients-et-proches/Les-cancers/Cancer-du-sein/Symptomes> (Consulté le 02/06/2021 à 21 :44).

- Des changements de forme de vos seins.

Un changement de la taille ou de la forme du sein

Une rougeur, un œdème (Un œdème désigne le gonflement d'un organe ou d'un tissu causé par l'accumulation d'un liquide séreux[1]) et une chaleur importante au niveau du sein peuvent être le signe d'un cancer du sein inflammatoire.

4. Les facteurs de risque

Les causes du cancer du sein sont multiples, à la fois d'origine génétique et environnementale. Mais, en dépit de très nombreuses études, elles ne sont pas entièrement connues. Les principaux facteurs de risque établis ou suspectés du cancer du sein sont :

Facteurs reproductifs

Âge aux premières règles

Le risque augmente avec la précocité de la survenue des premières règles. Chaque année de retard dans l'installation des premières règles s'associe à une réduction de 5 % du risque.[2]

Âge à la ménopause

Le risque augmente avec l'âge de la ménopause. Chaque année de retard dans l'installation de la ménopause s'associe à une augmentation de 3 à 4% du risque.[3]

Age à la première grossesse

[1] Le journal des femmes Santé, disponible sur : < https://sante.journaldesfemmes.fr/fiches-sante-du-quotidien/2514427-oedeme-causes-diagnostic-doppler-comment-le-resorber/>. (Consulté le 03/06/2021 à 09 :38).
[2] Jean-François Morère, Matti S. Aapro, Frédérique Penault-Llorca, Rémy Salmon, **Le cancer du sein**, Springer Paris, (2007), p 14.
[3] Ibid.

Même si le risque de cancer du sein apparait transitoirement augmenté pendant une durée de cinq à dix ans après la grossesse, il semble qu'il diminue lorsque la première grossesse survient avant 30 ans et est accru si elle survient après 35 ans.[1]

Parité

La parité, plutôt à un âge jeune, semble aussi avoir un effet protecteur. Chaque naissance réduit le risque de cancer du sein d'environ 7 %. Le rôle délétère de la nulliparité dans le cancer du sein est reconnu depuis longtemps (une épidémie de cancer du sein chez les nonnes a été décrite par le médecin Ramazzini à Padoue en 1743). Chez les femmes non nullipares, le risque de cancer du sein est d'autant plus faible que la parité est élevée.[2]

Allaitement

L'allaitement constitue un facteur protecteur du cancer du sein, Les femmes qui allaitent pendant au moins une année présentent un risque plus faible de cancer du sein que les autres, car l'allaitement peut apporter des changements dans les hormones et dans le tissu du sein qui favorisent la protection des cellules contre le cancer.[3]

Traitements hormonaux de la ménopause

Les traitements hormonaux de la ménopause sont associés à un risque accru de cancer du sein. Le risque est plus élevé chez les utilisatrices récentes de THM et persiste quelques années après l'arrêt du traitement, et il augmente avec la durée totale d'utilisation.[4]

[1] Ibid.
[2] Jean-Marc Classe, **Cancer du sein : Dépistage et prise en charge**, Elsevier Masson, (2016), p 6.
[3] Jean-François Morère, Matti S. Aapro, Frédérique Penault-Llorca, Rémy Salmon, Op.cit., p14.
[4] Jean-Marc Classe, Op.cit., p6.

Facteurs génétiques et démographiques

L'âge

L'âge est le facteur de risque le plus important du cancer du sein. L'incidence du cancer du sein augmente avec l'âge, doublant environ tous les dix ans jusqu'à la ménopause, période au cours de laquelle la courbe d'incidence tend à s'aplatir, en rapport avec l'arrêt de la production d'hormones stéroïdiennes par l'ovaire.[1]

Histoire familiale et mutations génétiques

L'histoire familiale est associée, de manière régulière, à un risque accru de cancer du sein. Le risque relatif pour toute forme de parenté est d'environ 1,9 et l'excès de risque est plus marqué chez les femmes plus jeunes et lorsque la maladie s'est développée chez une proche parente (mère, fille ou sœur), avant l'âge de 50ans.

Par ailleurs, certaines mutations génétiques sont susceptibles d'augmenter le risque de cancer du sein. Deux gènes, BRCA1 et BCRA2, semblent les plus impliqués. Par rapport à la population générale, les femmes porteuses des mutations sur ces gènes présentent un risque accru de cancer du sein.[2]

Facteurs liés aux habitudes de vie et nutrition

Obésité et prise de poids

L'obésité est associée à un profil hormonal soupçonné de favoriser le développement du cancer du sein. L'obésité augmente d'environ 50 % le risque de cancer du sein chez les femmes ménopausées. L'obésité n'augmente pas le risque chez les femmes avant la ménopause. Elle serait même associée à un risque réduit chez ces femmes dans les pays économiquement développés. Toutefois, l'obésité apparaît comme un facteur de risque important après la ménopause. Par ailleurs, les femmes ayant un surpoids de plus de 20 kg à partir de l'âge de 18

[1] Ibid., p 5.
[2] André Nkondjock, Parviz Ghadirian, **Facteurs de risque du cancer du sein**, M/S : médecine sciences, (2005), p 177-178.

ans, présentent, après la ménopause, un risque de cancer du sein multiplié par deux.[1]

Activité physique

L'activité physique modérée (30 à 60 minutes au moins 4 fois par semaine) diminue le risque de cancer du sein d'environ 35 %, en particulier chez les femmes ménopausées. L'activité physique influence également le risque de cancer du sein en diminuant la prise de poids, en particulier après la ménopause. L'obésité après la ménopause est un facteur de risque bien circonscrit et indépendant du cancer du sein ; elle peut être évitée par l'activité physique, une composante majeure du maintien de l'équilibre énergétique.[2]

[1] Ibid., p 178.
[2] Ibid., p 178-179.

Section 3 : Diagnostic, dépistage et traitement du cancer du sein

Comme tout autres types de cancer, le diagnostic joue un rôle très important dans le traitement du cancer du sein.

1. Dépistage du cancer du sein

Un dépistage consiste à détecter un cancer avant qu'il ne soit palpable ou qu'il ne se traduise par un signe anormal comme une modification de la peau ou du mamelon. L'examen utilisé pour dépister un cancer du sein est une mammographie (radiographie des seins), elle détecte des anomalies de petite taille, dont certaines seulement se révéleront être un cancer. Ces anomalies sont parfois détectées même si l'examen clinique est normal.[1]

2. Diagnostic du cancer du sein

Comment découvre-t-on un cancer du sein ?

Un cancer du sein est le plus souvent diagnostiqué à quatre occasions :[2]

● **Lors de la découverte de symptômes par la patiente elle-même.**

● **Lors d'une consultation de dépistage.**

● **Lors d'une consultation habituelle chez le gynécologue :** Lors d'une visite de contrôle, le médecin peut trouver une anomalie au niveau des seins

● **Lors de la surveillance d'un premier cancer du sein :** Lors de la surveillance d'un cancer du sein traité, le médecin vérifie qu'un second cancer du sein ne s'est pas développé.

[1] Institut national du cancer, **guide d'information : Comprendre le cancer du sein**, (2007), p 19.
[2] Ibid., p 19.

3. Traitement du cancer du sein

Il existe plusieurs types de traitements pour le cancer du sein comme toute autres types de cancer et chaque type de traitement est choisi selon :[1]

L'objectif :

• Supprimer la tumeur ou les métastases ;

• Réduire le risque de récidive ;

• Ralentir le développement de la tumeur ou des métastases ;

• Améliorer le confort et la qualité de vie de la personne malade, en traitant les symptômes engendrés par la maladie.

Et selon les caractéristiques suivantes :

• Du type de cancer dont vous êtes atteinte 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 ;

• De votre état de santé général, de votre âge, de vos antécédents personnels médicaux et chirurgicaux et de vos antécédents familiaux ;

• De votre avis et de vos préférences.

Les types de traitements du cancer du sein

Différents types de traitements peuvent être utilisés pour traiter un cancer du sein : la chirurgie, la radiothérapie, la chimiothérapie, et l'hormonothérapie.

[1] Institut national du cancer (France), patient et proche, traitement, disponible sur : <https://www.e-cancer.fr/Patients-et-proches/Les-cancers/Cancer-du-sein/Traitements>. (Consulté le 03/06/2021 à 10 :23).

La chirurgie

La chirurgie est le traitement le plus anciennement utilisé pour soigner les cancers du sein. C'est un traitement standard. Et la chirurgie du cancer du sein a quatre objectifs : [1]

• confirmer le diagnostic et préciser le stade d'évolution du cancer, notamment examiner si les ganglions ont été atteints par des cellules cancéreuses ;

• enlever la tumeur ;

• prélever et examiner certains ganglions ;

• conserver ou restaurer la taille et la forme du sein après l'ablation de la tumeur ou de la totalité du sein.

La radiothérapie

La radiothérapie est utilisée depuis de longues années pour traiter différents cancers. Elle consiste à recourir à des rayons X qui atteignent la tumeur et détruisent les cellules cancéreuses. Selon la zone et les organes de voisinage à traiter, les rayons utilisés sont différents (photons ou électrons). Ces divers types de rayons sont parfois associés entre eux. C'est pour cette raison que, pour un même traitement, une patiente peut être placée sous différents appareils : appareils de cobalthérapie ou accélérateurs linéaires.[2]

La chimiothérapie

La chimiothérapie est l'un des traitements du cancer du sein. Il s'agit d'un traitement dit adjuvant, c'est-à-dire qui complète un traitement chirurgical. La chimiothérapie adjuvante a pour objectif de diminuer le risque que des cellules cancéreuses se développent à distance, ce qu'on appelle des métastases.
La chimiothérapie agit sur le cancer à l'aide de médicaments appelés médicaments antitumoraux ou médicaments anticancéreux. Ces médicaments agissent par voie

[1] Institut national du cancer, **guide d'information : Comprendre le cancer du sein**, (2007), p 34.
[2] Ibid., p 48.

générale : ils agissent sur les cellules cancéreuses dans l'ensemble du corps, soit en les détruisant, soit en les empêchant de se multiplier.[1]

L'hormonothérapie

L'hormonothérapie est l'un des traitements du cancer du sein. C'est un traitement général qui agit dans l'ensemble du corps. Certaines hormones secrétées par les ovaires* stimulent la croissance des cellules cancéreuses. Une hormonothérapie vise à empêcher l'action de ces hormones ou à diminuer leur sécrétion afin de ralentir ou de stopper la croissance des cellules cancéreuses. Il s'agit d'un traitement adjuvant qui complète le traitement local.[2]

[1] Ibid., p 51.
[2] Ibid., p 61.

Conclusion

Ce premier chapitre « Généralités sur le cancer du sein. » nous a permis de :

› Comprendre de manière générale le cancer et de bien comprendre le cancer du sein et de voir qu'il est le cancer le plus fréquent chez la femme dans le monde et en Algérie.

› Connaitre les différents types de cancer du sein et ses facteurs les plus importants, et qui sont utilisé dans notre étude.

› Voir les différents types de traitements et connaitre comment le cancer du sein est diagnostiqué et comprendre le dépistage.

CHAPITRE 02 : Les fondements théoriques de Data mining et ses techniques

Introduction

Le but de ce travail est de construire un modèle de classification et de prédiction qui permet de prédire la classe des femmes soit Malade ou bien Non malade, et aussi de voir quel facteur de risque parmi les facteurs cités dans le chapitre précédent explique le plus le cancer du sein selon notre échantillon.

Pour atteindre notre objectif, nous utiliserons des outils ou bien des techniques du Data mining, et pour bien comprendre ses techniques là et de quoi consiste le Data mining, nous avons partagé ce chapitre en trois sections :

« Section 1 : Introduction au Data mining et au Knowledge discovery in databases » : dans cette section nous avons défini le Data mining et ses techniques, et aussi le KDD et son processus, et cette section nous a permet de connaitre le lien entre ces deux termes (Le KDD et le DM).

Dans la deuxième section « Les arbres de décision », nous avons vu les arbres de décision et leurs types, et comment ils sont construits, et connaitre ses avantages et inconvénients.

Enfin nous avons brièvement défini le classifieur bayésien naïf et ses types dans la dernière section « La classification bayésienne naïve ».

Section 1: Introduction au Data mining et au Knowledge discovery in databases

L'augmentation rapide des données stockées dans les bases de données a conduit à l'émergence d'un domaine appelé Data mining et découverte de connaissances (KDD) afin d'extraire les connaissances cachées dans ces données pour améliorer le processus de prise de décision de l'entreprise ou l'organisation. Et dans ce chapitre nous expliquons ces deux termes le DM et le KDD.

1. Historique

L'expression « Data mining » est apparue vers le début des années 1960 et avait, à cette époque, un sens péjoratif. En effet, les ordinateurs étaient de plus en plus utilisés pour toutes sortes de calculs qu'il n'était pas envisageable d'effectuer manuellement jusque-là. Certains chercheurs ont commencé à traiter sans a priori statistique les tableaux de données relatifs à des enquêtes ou des expériences dont ils disposaient. Comme ils constataient que les résultats obtenus, loin d'être aberrants, étaient encourageants, ils furent incités à systématiser cette approche opportuniste. Les statisticiens officiels considéraient toutefois cette démarche comme peu scientifique et utilisèrent alors les termes « Data mining » ou « Data Fishing » pour les critiquer.

L'analyse des données s'est développée et son intérêt grandissait en même temps que la taille des bases de données. Vers la fin des années 1980, des chercheurs en base de données, tel que Rakesh Agrawal, ont commencé à travailler sur l'exploitation du contenu des bases de données volumineuses comme par exemple celles des tickets de caisse de grandes surfaces, convaincus de pouvoir valoriser ces masses de données dormantes. Ils utilisèrent l'expression « Database Mining » mais, celle-ci étant déjà déposée par une entreprise (Database Mining Workstation), ce fut « Data mining » qui s'imposa. En mars 1989, Shapiro Piatetski proposa le terme « Knowledge Discovery » à l'occasion d'un atelier sur la découverte des connaissances dans les bases de données. Actuellement, les termes Data mining et Knowledge Discovery in data bases (KDD, ou ECD en français) sont utilisés plus ou moins indifféremment. Nous emploierons par

conséquent l'expression « Data mining », celle-ci étant la plus fréquemment employée dans la littérature.

La communauté de « Data mining » a initié sa première conférence en 1995 à la suite de nombreux workshops sur le KDD entre 1989 et 1994. En 1998, s'est créé, sous les auspices de l'ACM, un chapitre spécial baptisé ACMSIGKDD, qui réunit la communauté internationale du KDD. La première revue du domaine « Data mining and Knowledge Discovery journal » publiée par « Kluwers » a été lancée en 1997.[1]

2. Qu'est-ce que le Data mining et le KDD

Définition du Data mining

Le Data mining ou exploration de données en français, est un ensemble de techniques statistiques et d'apprentissage machine qui sont misent en œuvre afin d'extraire et de rechercher l'information pertinentes cachées dans les données, pour l'aide à la décision et à la prévision. [2]

Et le Data mining a été défini par Sholom M. Weiss & Nitin Indurkhya comme suit :

"L'exploration de données est la recherche d'informations précieuses dans de grands volumes de données. Il s'agit d'un effort de coopération entre les humains et les ordinateurs. Les humains conçoivent des bases de données, décrivent des problèmes et fixent des objectifs. Les ordinateurs passent les données au crible, à la recherche de modèles qui correspondent à ces objectifs."[3]

[1] Djamel Abdelkader ZIGHED & Ricco RAKOTOMALALA, Extraction des Connaissances à partir des Données (ECD), Data mining, p 5-6.
[2] LOUNICI MOSBAH.N, Data mining et apprentissage, Chapitre I: introduction au Data mining, p 7.
[3] Sholom M. Weiss & Nitin Indurkhya, Predictive Data Mining: A Practical Guide, First Edition, p 1.

Définition du Knowledge discovery in databases

Le terme Knowledge discovery in databases (KDD) désigne la procédure générale de l'extraction et la découverte de connaissances dans les données. Et le KDD a été défini par Fayyad, Piatetsky-Shapiro et Smyth en 1996 comme suit :

"Knowledge Discovery in Databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. "[1]

Le processus KDD est une démarche qui consiste à utiliser une base de données avec toute sélection, prétraitement, sous-échantillonnage et transformation de données requise, et d'appliquer les algorithmes du Data mining pour construire des modèles et les évaluer et de les mettre en œuvre afin d'extraire des connaissances.[2]

Le Data mining et le KDD

Les termes Knowledge discovery in databases et Data mining sont utilisés de manière interchangeable, ils font référence à deux concepts liés mais légèrement différents. Le KDD est le processus global d'extraction de connaissances à partir de données, et le Data mining en étant un ensemble de techniques et d'algorithmes a été inclus dans le processus KDD comme sous-partie (étape) consacré à l'application d'algorithmes capables de construire des modèles à partir des données. Et ses deux termes aujourd'hui sont souvent considérés identiques.[3]

[1] Usama Fayyad, Gregory Piatetsky-Shapiro & Padhraic Smyth, Knowledge Discovery and Data Mining: Towards a Unifying Framework, 1996, p 83.
[2] Ibid.
[3] Fathalrahman Adam & Fathalrahman Adam, Knowledge Discovery in Big Data from Astronomy and Earth Observation: AstroGeoInformatics, 2020, p 1.

3. Les taches et méthodes du Data mining

Les taches du Data mining

Plusieurs taches peuvent être associées au Data mining qui se résument, brièvement, comme suit :[1]

La classification

Elle consiste à examiner les caractéristiques d'un objet et lui attribuer une classe (p. ex., décision d'attribution de prêt à un client).

La prédiction

Prédire la valeur future d'un attribut en fonction d'autres attributs (p. ex., prédire la qualité d'un client).

L'estimation

Elle consiste à estimer la valeur d'un champ (avec une variable cible qui est numérique) à partir des caractéristiques d'un objet.

L'association

Est la tâche la plus intéressante du Data mining et la plus répandue dans le monde des affaires, elle consiste à chercher des règles d'association et à découvrir les règles de quantification ou de relation entre deux ou plusieurs attributs. C-à-d elle consiste à déterminer les attributs qui sont corrélés (associés) (p. ex., analyse du panier de la ménagère).

La description

Elle nous sert à expliquer ou vérifier un fait, et cela en essayant de trouver des façons de décrire les tendances cachées dans les données.

[1] BOUDHEB Tarik, Thèse de doctorat: Privacy Preserving Classification of Biomedical Data, Université Djilali Liabes Sidi BEL ABBES, 2018/2019, p 32.

La segmentation (Clustering)

Elle consiste à former des groupes homogènes à l'intérieur d'une population. Et dans ce cas-là il n'y a pas de variable cible.

Les méthodes du Data mining

Il existe plusieurs méthodes différentes utilisées pour effectuer les tâches du Data mining et ces techniques font appel à différents algorithmes pour accomplir ces différentes tâches, et tous ces algorithmes tentent d'ajuster un modèle aux données. Et ces techniques-là sont classées en deux catégories :[1]

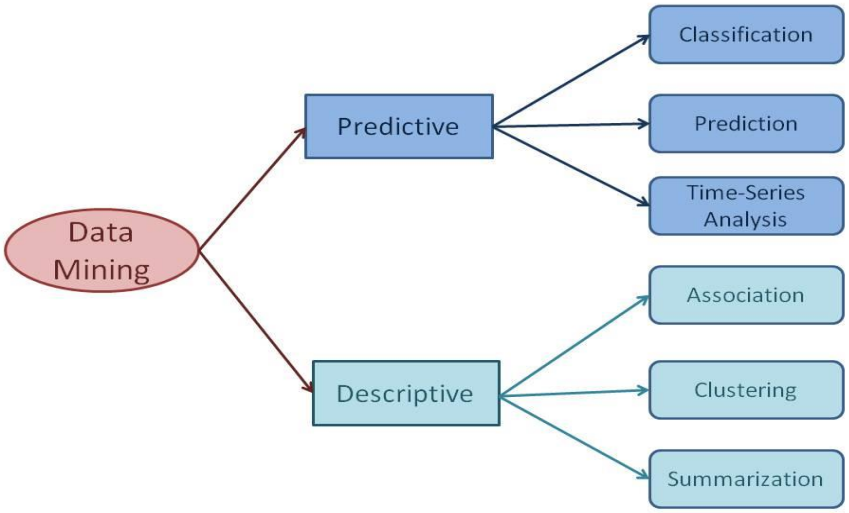

Figure (II-9) : Classification des techniques du Data mining.

Source : WideSkills.

Les techniques prédictives

Ces techniques emploient des algorithmes dont l'objectif est de construire un modèle basé sur un ensemble d'observations, afin de prédire la valeur ou bien la classe d'une nouvelle variable. En d'autres termes, ces algorithmes examinent les

[1] Massinissa Saoudi, Thèse de doctorat : Conception of a wireless sensor network for decision making based on Data mining methods, Université de Bretagne occidentale - Brest, 2017, p 24.

données historiques pour prédire l'avenir d'une variable dont ont souhaité connaitre sa valeur ou sa classe qu'on l'appelle variable cible.[1]

Les techniques descriptives

L'objectif de ces techniques est de caractériser les propriétés générales de données et de trouver un modèle. En d'autres termes, ils décrivent les données de manière concise et présentent leurs propriétés intéressantes.[2]

4. Domaines d'application du Data mining

Le Data mining est applicable dans plusieurs domaines et secteur parmi eux :

La santé

L'exploration de données peut améliorer différent aspect du système de santé, car elle utilise à la fois l'analyse et les données pour obtenir de meilleures informations à partir des bases de données, et le secteur de santé peut utiliser ces informations pour déterminer les bonnes solutions afin d'améliorer les services de santé et réduire les couts.[3]

L'éducation

L'utilisation du Data mining dans l'éducation est relativement récente et l'objectif de cette utilisation dans ce secteur est d'explorer des connaissances à partir de grands volumes de données provenant d'environnement éducatif, afin de comprendre le comportement futur des étudiants et pour prendre les bonnes décisions pour aider les étudiants à s'améliorer, et la fouille de données est aussi utilisée pour prédire les notes des étudiants.[4]

[1] Ibid.
[2] Ibid.
[3] Alex Campbell, Data Visualization Guide: Clear Introduction to Data mining, Analysis, and Visualization, (2021), p 7.
[4] Ibid., p 7.

L’analyse du panier de la ménagère

Cette forme d’analyse repose sur différentes hypothèses. Si vous achetez des produits spécifiques, vous allez probablement acheter un autre produit du même groupe. Cette forme d’analyse permet au vendeur ou détaillent d’identifier plus facilement le comportement d’achat de ses clients et de connaitre leurs besoins, ce qui lui permet de modifier plus facilement l’aménagement du magasin.[1]

La gestion de la relation client (GRC)

L’objectif de la GRC est d’acquérir et de maintenir des clients, et permet aux entreprises de développer des stratégies axées sur le client et d’améliorer la fidélité de la clientèle.[2]

Les services bancaires financiers

Les banques ont pris un tournant et ont commencé à numériser et stocker toutes les transactions et les informations liées aux clients. Et en utilisant des algorithmes et techniques du Data mining, les banquiers peuvent résoudre divers problèmes liés aux activités de la banque. Et ils peuvent utiliser ces techniques et méthodes lorsqu’ils travaillent avec de gros volumes de données, et ces techniques-là facilitent aux gestionnaires et aux experts la tâche d’acquérir, cibler, segmenter, maintenir et conserver divers profils de clients.[3]

5. Processus de Data mining

Le processus de la fouille de données, ou le KDD (Knowledge Data Discovery), comprend les étapes suivantes :

[1] Ibid.
[2] Ibid., p 8.
[3] Ibid., p 9.

Figure (II-10) : Les étapes d'une étude de Data mining.

Source : Elaboré par l'étudiant à l'aide de logiciel Word2019.

Définition des objectifs

Tout d'abord, il faut commencer par :[1]

• Choisir le sujet et les objectifs à atteindre ou problèmes à résoudre.

• Définir la population cible (les prospects et les clients, seulement les clients, seulement les clients fidèles, tous les malades, seulement les malades curables par le traitement teste…).

• Définir l'entité statistique étudiée et définir certains critères essentiels et en particulier le phénomène à prédire, planifier le projet, prévoir l'utilisation opérationnelle des informations extraites et des modèles produits, et spécifier les résultats attendus.

• Cette étape conditionne en partie d'avoir une idée sur la catégorie de méthodes du Data mining (soit prédictives ou descriptives), et de choisir une ou plusieurs techniques qu'on souhaitera utiliser pour atteindre nos objectifs.

Et le choix du type de modèle influencera la préparation de données qu'on doit effectuer et la façon dont on y prendra.

[1] Stéphane Tufféry, Data mining et statistique décisionnelle : La science des données, 5ème édition, p 33-34.

Collecte des données

Cette étape-là est absolument essentielle, c'est au cours de cette étape que notre base de données est construite, et la collecte des données passe par les deux phases suivantes :[1]

La première concerne le recensement des données utiles, accessibles (internes ou externes à l'entreprise ou à l'organisation), légalement et techniquement exploitables, fiables et suffisamment à jour. Ces données proviennent de l'intérieur de l'entreprise (stockées dans le système d'information ou bien hors système d'information), ou sont achetées ou récupérées à l'extérieur de l'entreprise ou l'organisation.

Et dans cette deuxième phase qu'on constitue la base de données qui servira à la construction des modèles. Cette base d'analyse se présente le plus souvent sous la forme d'une table (Oracle, MySQL, Microsoft SQL Server…) ou d'un fichier (à plat, CSV…), ayant un enregistrement (une ligne) par individu statistique étudie et un champ (une colonne) par variable de cet individu.

Exploration et préparation des données

Et là on effectue un ensemble d'opération pour explorer et préparer les données pour les utiliser dans l'étape de la construction, et ces opérations sont les suivantes :[2]

Data cleaning

Représente les opérations de base qui éliminent le bruit ou les données incohérentes, la réduction des données et le traitement des données manquantes.

Data integration

Est utilisé lorsque plusieurs sources de données peuvent être combiné.

[1] Ibid., p 35.
[2] Massinissa Saoudi, Op.cit., p 22-23.

Data selection

Est utilisé pour récupérer des données pertinentes à l'analyse à partir de la base de données.

Data transformation

Est utilisé pour transformer ou consolider des données dans les formulaires appropriés en effectuant des opérations de synthèse ou d'agrégation.

Construction des modèles

L'étape de construction de modèles est considérée comme le cœur du processus d'extraction de connaissances à partir des données. Elle consiste à appliquer les techniques qui ont été choisi dans la première étape sur les données collectées et traitées dans l'étape 2 et 3 afin de construire un modèle.

La construction des modèles prédictifs nécessite un protocole d'entrainement et de validation bien défini afin d'assurer que les prédictions soient les plus précises et les plus robustes. Ce type de protocole est parfois appelé l'apprentissage supervisé, il consiste à partager la base de données en deux parties, la première est appelée *Training set*, il sert à estimer le modèle, puis à la tester et la valider sur la deuxième partie *Testing set*.[1]

Evaluation des résultats obtenues et mise en œuvre des connaissances

Cette étape nous permet d'évaluer et interpréter les modèles obtenus dans l'étape précédente et de comparer entre eux et identifier les modèles les plus intéressants. Et ensuite la mise en œuvre des connaissances dans un autre système pour une activité ultérieure qui seront utilisées pour comprendre un système et prédire les actions.[2]

[1] Herbert A. Edelstein, Introduction to Data Mining and Knowledge Discovery, Third Edition, p 27-28.
[2] Massinissa Saoudi, Op.cit., p 23.

Section 2 : Les arbres de décision

Les arbres de décision sont parmi les méthodes les plus utilisés en Data mining, elles sont applicables pour tout type de problème et tout type de données.

1. Définition de l'arbre de décision

Dans la théorie des graphes, un arbre est un graphe connexe et acyclique, où connexe et acyclique veulent dire lié et sans cycle. Et sa forme évoque en effet la ramification des branches d'un arbre.[1]

Les arbres de décision (AD) sont une catégorie d'arbres utilisée dans le Data mining et en informatique décisionnelle, ils emploient une représentation hiérarchique de la structure des données sous forme des séquences de décisions (tests) en vue de la prédiction d'un résultat ou d'une classe. Chaque individu (ou observation), qui doit être attribué(e) à une classe, est décrit(e) par un ensemble de variables qui sont testées dans les nœuds de l'arbre. Les tests s'effectuent dans les nœuds internes et les décisions sont prises dans les nœuds feuilles.[2]

Un arbre de décision se compose de trois sortes de nœuds, et des branches :

Nœud racine : l'accès à l'arbre se fait par ce nœud.

Nœuds internes (ou tests) : ce sont des nœuds qui ont des descendants (ou enfants), qui sont à leurs tour des nœuds.

Nœuds terminaux (ou feuilles/classes) : nœuds qui n'ont pas de descendant, et si tous les individus appartiennent à la même classe, le nœud terminal est déclaré pur.

Branches : représentent les alternatives de chaque test qui nous permet de se diriger vers un autre nœud fils.

[1] Jean-Claude Fournier, Théorie des graphes et applications : avec exercices et problèmes, 2ème édition revue et augmentée, p 41.
[2] Le CNAM, cours : Arbres de décision, disponible sur : <https://cedric.cnam.fr/vertigo/Cours/ml2/coursArbresDecision.html> (Consulté le 09/08/2021 à 01 :18).

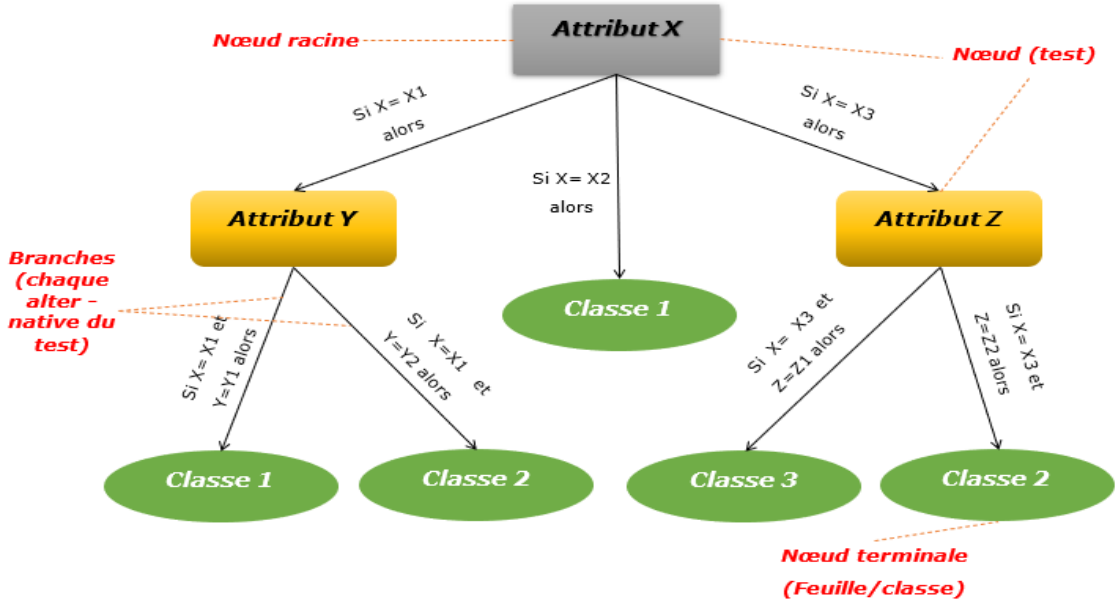

Figure (II-11) : Les composantes d'un arbre de décision.

Source : Elaboré par l'étudiant à l'aide de logiciel Word2019.

Pour parcourir un arbre de décision et trouver une solution il faut passer par les étapes suivantes :

• Choisir l'instance qu'on veut tester ou classer et débuter à la racine de l'arbre.

• Descendre dans l'arbre en passant par les nœuds internes (nœuds tests), et chaque nœud est une décision atomique. Chaque réponse possible et prise en compte permet de se diriger vers des fils du nœud.

• Et de proche en proche en descendant dans l'arbre, on atteint à la fin une feuille qui nous permet de classer l'instance testée.

2. Les principaux arbres de décision

Les principaux algorithmes d'arbres de décision sont :

CART (Classification And Regression Tree)

L'arbre CART, inventé en 1984 par les statisticiens L. Breiman, J.H. Friedman, R.A. Olshen et C.J. Stone, est l'un des arbres de décision les plus efficaces et les plus répandus. CART recourt à l'indice de Gini pour trouver la

meilleure division de chaque nœud, et prend à chaque fois l'indice de Gini le plus petit.[1]

C5.0

L'arbre C5.0 est un perfectionnement par le chercheur australien J. Ross Quinlan de ses précédents arbres ID3 et C4.5. Le C5.0 est moins répandu que CART et fonctionne en cherchant à maximiser le gain d'information réalisé en affectant chaque individu à une branche de l'arbre. Et le C5.0 partage avec le CART sont adaptation à l'étude de tout type de variables, et sa recherche exhaustive de toutes les scissions possibles ainsi que son dispositif d'optimisation de l'arbre par construction puis élagage d'un arbre maximum.[2]

CHAID (Chi-squared Automatic Interaction Detector)

Le CHAID est une technique plus ancienne, son principe remonte à 1975 et son algorithme à 1980. Il utilise le test du χ^2 pour définir la variable la plus significative de chaque nœud, il ne peut être utilisé qu'avec des variables explicatives discrètes ou qualitatives.[3]

3. Comment construire un arbre de décision ?

Pour construire un arbre de décision faut passer par les étapes suivantes :

Le choix de la variable de segmentation

Avant de construire un arbre de décision, on commence par choisir une variable (ou attribut) qui sépare le mieux les individus de chaque classe, en fonction des valeurs de cette variables.
Et la variable de segmentation est choisi selon plusieurs critères qui sont :[4]

Le critère de X^2 Khi-deux

[1] Stéphane Tufféry, Op.cit., p 661-662.
[2] Ibid., p 665-666.
[3] Ibid., p 667.
[4] Ibid., p 653.

Le critère Khi-deux est utilisé lorsque les variables explicatives sont qualitatives ou bien discrètes (utilisé dans l'arbre de CHAID).

Le critère de Gini

Ce critère est utilisable pour tout type de variables explicatives et il est utilisé dans l'arbre CART.

L'entropie, ou information

Est utilisé pour tout type de variables explicatives (utilisé dans les arbres CART, C4.5, et C5.0).

Le critère Twoing

Ce critère est utilisé pour tout type de variables explicatives, et lorsque la variable à expliquer (variable d'intérêt ou cible) a k≥3 modalités et que l'on veut ramener à la recherche d'une scission optimale sur k modalités à une scission optimale sur 2 super-modalités composées de modalités initiales.

Le critère Twoing ordonné

Lorsque la variable à expliquer à k≥3 modalités ordonnées. Critère dans lequel les deux super-modalités ne regroupent que des modalités adjacentes parmi les modalités initiales.

Le traitement des variables continues

Pour le traitement des variables continues faut qu'on détermine le meilleur point de coupure ou le seuil de découpage le plus pertinent. L'idée est de transformer la variable quantitative en variable booléenne (X≤ ou >) et pour cela il faut :[1]

[1] LOUNICI MOSBAH.N, Data-Mining et apprentissage, Chapitre II : Modèle des arbres de décision, p 46-47.

• Classer les valeurs du *Training set* par ordre croissant et les annotées de leurs classes.

• Calculer les seuils entre chaque deux variables : *Seuil= (X_i+X_{i+1}) /2.*

• Choisir le point de coupure qui minimise le critère de Gini ou maximise le gain d'information (selon le type d'arbre).

La définition de la bonne taille de l'arbre

Les performances d'un arbre de décision reposaient principalement sur la détermination de sa taille. Les arbres ont tendances à produire un classifieur trop complexe, collant exagérément aux données ; c'est le phénomène de « sur-apprentissage ». Les feuilles, mêmes si elles sont pures, sont composées de trop peu d'individus pour être fiables lors de la prédiction. Et la taille de l'arbre a tendance à croitre avec le nombre d'observations dans la base d'apprentissage. Et l'enjeu de la recherche de la taille optimale consiste à stopper (pré-élagage) ou à réduire (post-élagage) l'arbre de manière à obtenir un bon classifieur.[1]

Pré-élagage

Le pré-élagage consiste à fixer une règle d'arrêt qui permet de stopper la construction de l'arbre lors de la phase de construction. Une approche très simple consiste à fixer un critère d'arrêt local, relatif au sommet que l'on est en train de traiter, qui permet d'évaluer l'apport informationnel de la segmentation que l'on va initier.[2]

Post-élagage

Le principe est de construire l'arbre en deux temps : une première phase d'expansion, où l'on essaie de produire des arbres les plus purs possibles et dans laquelle nous acceptons toutes les segmentations même si elles ne sont pas pertinentes ; dans un second temps, nous essayons de réduire l'arbre en utilisant un autre critère pour comparer des arbres de tailles différentes. Le temps de

[1] Ricco RAKOTOMALALA, Arbres de décision, Revue MODULAD, 2005, p 171.
[2] Ibid., p 172.

construction de l'arbre est bien sûr plus élevé ; il peut être pénalisant lorsque la base de données est de très grande taille ; en contrepartie, l'objectif est d'obtenir un arbre plus performant en classement.[1]

Décision

Cette étape consiste à affecter une conclusion à chaque feuille de l'arbre. Le chemin reliant une feuille à la racine de l'arbre peut être lu comme une règle de prédiction du type attribut-valeur (Si prémisse… alors Conclusion…). En revanche, lorsque plusieurs modalités sont présentes dans la feuille, il faut utiliser une règle d'attribution efficace. La règle la plus souvent utilisée est la règle de la majorité : on affecte à la feuille la modalité de la variable à prédire qui présente l'effectif le plus grand.[2]

Evaluation d'un modèle de prédiction

Matrice de confusion

Une manière classique d'évaluer la qualité de l'apprentissage est de confronter la prédiction du modèle avec les valeurs observées sur un échantillon de la population. Cette confrontation est résumée dans un tableau croisé appelé matrice de confusion.

Valeur actuelle	Positive	Vrai positif	Faux positif
	Négative	Faux négatif	Vrai négatif
		Positive	Négative
		Valeur prédite	

Figure (II-12) : Matrice de confusion pour une classification binaire.

Source : Elaboré par l'étudiant à l'aide de logiciel Word2019.

[1] Ibid., p 173.
[2] Ibid., 174.

Indicateurs de la performance

Ces indicateurs sont calculés à partir de la matrice de confusion et on trouve :

VP : Vrai positif ; FP : Faux positif ; VN : Vrai négatif ; FN : Faux négatif.

P : nombre des positives, avec P= VP+FP.

N : Nombre des négatives, avec N= FN+VN.

Taux d'erreur=$\frac{FP+FN}{N+P}$ **,** c'est la proportion d'individus (observations) mal classés par notre modèle.

Sensibilité=$\frac{VP}{VP+FN}$ **,** il mesure la capacité du modèle à détecter les vrais positives.

Spécificité=$\frac{VN}{VN+FP}$ **,** il mesure la capacité du modèle à détecter les vraies négatives.

Précision=$\frac{VP+VN}{N+P}$ **,** la précision mesure la proportion d'individus que notre modèle a prédit correctement.

4. Les avantages et inconvénients

Les avantages et inconvénients ont été cité par Stéphane Tufféry comme suit :[1]

Les avantages

• Les arbres de décision sont simples à utiliser et les résultats sont exprimés sous formes de conditions explicites sur les variables d'origines.

[1] Stéphane Tufféry, Op.cit., p 672-676.

• La technique des arbres de décision est non-paramétrique, c'est-à-dire qu'elle ne suppose pas que les variables explicatives suivent des lois probabilistes particulières.

• L'arbre sont peu perturbés par la présence d'individus hors norme, et les arbres peuvent gérer les données manquantes.

• Certains types permettent de traiter directement tous types de variables (continues, discrètes et qualitatives) comme le CART et C4.5.

Les inconvénients

• La définition des nœuds au niveau *n+1* dépend fortement de celle au niveau *n*.

• L'apprentissage d'un arbre de décision nécessite un nombre suffisamment grand d'individus au moins une trentaine par nœud pour que les règles aient un sens.

• On a une discontinuité de la réponse de la variable à expliquer en fonction des variables explicatives. Un petit changement de la valeur de X peut valider ou invalider une règle, et modifier complètement la prédiction de l'arbre.

• Temps de calcul important pour la recherche des critères de division et élagage.

Section 3 : La classification bayésienne naïve

La classification comme nous l'avons vu dans ce chapitre fait référence à la tâche de prédiction de classe d'une observation, et dans ce chapitre nous examinons l'une des méthodes de classification probabiliste qui est la classification bayésienne naïve.

Le théorème de Bayes

Le théorème de bayes porte le nom de Thomas Bayes (1702-1761), un statisticien et philosophe anglais, et ce théorème est l'un des plus importants de la théorie des probabilités. Il s'agit d'un théorème d'inversion des probabilités qui relie pour deux événements *A* et *B* la probabilité conditionnelle de *A* sachant *B* à la probabilité conditionnelle de *B* sachant *A* :[1]

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

Définition de la classification bayésienne naïve

La classification bayésienne naïve est une méthode de classification qui met en œuvre un classifieur bayésien naïf appartenant à la famille des classifieurs linéaires, et cet indicateur est composé de deux mots bayésien et naïf, le premier fait référence au théorème de bayes et le deuxième mot « naïf » est employé car le classifieur suppose que toutes les variables sont indépendantes les unes des autres. Ce qui est atypique des exemples du monde réel, mais malgré cette hypothèse d'indépendance, le classifieur bayésien naïf donne souvent de bons résultats dans beaucoup de secteurs et domaine comme le secteur de santé.[2]

[1] Stéphane Tufféry, Op.cit., p 679.
[2] Max Bramer, Principles of Data Mining, Fourth Edition, 2020, p 22.

Principe

Le principe du classifieur bayésien naïf est de calculer les probabilités conditionnelles d'appartenance d'un individu I_i ayant des caractéristique X= {x_1, x_2…, x_p} (variables explicatives) à une classe C_i parmi C_k classes, en tenant compte l'hypothèse d'indépendance des variables explicatives X avec la variable d'intérêt, et de retenir la classe C qui maximise cette probabilité, pour attribuer notre individu à cette classe.[1]

Et la probabilité P (C_i / X) est calculé comme suit :

$$P(C_i/X) = \frac{P(C_i)P(X = x_1, x_2, \ldots . x_p/C_i)}{P(X = x_1, x_2, \ldots . x_p)}$$

En pratique, seul le numérateur nous intéresse, puisque le dénominateur ne dépend pas de C et les valeurs des caractéristiques X_i sont données. Le dénominateur est donc en réalité constant. Le numérateur est soumis à la loi de probabilité à plusieurs variables.

$$P(C_i/X) = P(C_i)P(X = x_1, x_2, \ldots . x_p/C_i)$$

Et vue que les p variables explicatives X= {x_1, x_2…, x_p} sont supposées indépendants, la probabilité de la combinaison est le produit des probabilités.

$$P(X = x_1, x_2, \ldots . x_p/C_i) = \prod_{j=1}^{p} P(x_j/C_i)$$

$$P(C_i/X = x_1, x_2, \ldots . x_p) = \prod_{j=1}^{p} P(x_j/C_i) * P(C_i)$$

[1] Stéphane Tufféry, Op.cit., p 680.

Et $P(C_i)$ est estimée par la fréquence relative :

$$\hat{P} = \frac{n_i}{n}$$

Avec n_i le nombre d'individu appartenant à C_i , et n le nombre d'observation total.

Les types de classifieur bayésien naïf

Comme ça été cité dans le principe du classifieur bayésien naïf le numérateur est soumis à la loi de probabilité à plusieurs variables, et cette loi est estimée selon le type des valeurs de X (les variables explicatives) :[1]

Loi normale

Lorsque les valeurs des caractéristiques sont continues, on utilise la loi normale (loi gaussienne). Par exemple, le poids, le prix, etc. En se basant sur les données d'entrainement avec N échantillons, on calcule l'espérance μ et la variance σ^2 de chaque caractéristique X et chaque classe C.

Loi multinomiale

Lorsque les valeurs des caractéristiques sont discrètes, on utilise la loi multinomiale. Par exemple, la couleur des cheveux avec les valeurs : brun, auburn, châtain, roux, blond vénitien, blond et blanc.

Loi de Bernoulli

Lorsque les valeurs des caractéristiques sont binaires, on utilise la loi de Bernoulli.

[1] Github, Introduction à l'apprentissage automatique, disponible sur : <https://projeduc.github.io/intro_apprentissage_automatique/bayes.html> (Consulté le 16/08/2021 à 01 :48).

Avantages et inconvénients

Le classifieur de Bayes présente plusieurs avantages et inconvénients :

• La classification bayésienne naïve est un processus informatique simple, facile à construire même pour un grand ensemble de données d'apprentissage.[1]

• Très utilisée par les chercheurs et donne de bons résultats.

• En théorie, le classifieur bayésien a le taux d'erreur minimum par rapport aux autres classificateurs. Par contre en pratique ce n'est pas toujours le cas en raison d'imprécisions dans les hypothèses d'attributs et d'indépendance conditionnelle de classe.[2]

• L'existence d'une grande corrélation entre les variables explicatives influencent les résultats obtenus et la performance du modèle.

[1] Mehmed Kantardzic, DATA MINING : Concepts, Models, Methods, and Algorithms, Third Edition, p 173.
[2] Ibid., p 174.

Conclusion

Le deuxième chapitre « Les fondements théoriques de Data mining et ses techniques » nous a permis dans la première section de faire une introduction sur le Data mining et de connaitre le processus du Knowledge Discovery in Databases et de faire le lien entre ces deux termes (le KDD et le DM), et de voir les différentes tâches et techniques du Data mining.

Et dans les deux dernières sections, nous avons vu deux techniques parmi les techniques de classification de Data mining et qui seront employées pour atteindre nos objectifs, et ces deux techniques sont la classification par arbres de décision et la classification bayésienne naïve, ce qui nous a permet de connaitre le principe de chaque technique et les différentes étapes de construction de chaque modèle, et connaitre leurs avantages et inconvénients.

CHAPITRE 03 : L’application des techniques du Data mining et interprétation des résultats

Introduction

L'objectif de ce travail est de collecter des informations sur des femmes atteintes d'un cancer du sein et les femmes non malades afin de voir lequel de ces facteurs est le plus discriminants et explique mieux cette maladie, et de construire à partir de ces facteurs un modèle de classification qui permet de prédire la classe de chaque femme (Malade ou bien Non malade).

Et pour cela nous avons partagé ce chapitre en trois sections afin de construire nos modèles :

La première porte sur la présentation de notre enquête en ligne et la présentation de la base de données obtenues, et englobe l'ensemble des opérations de traitement de données manquantes ou aberrantes afin d'obtenir un tableau de données prêt à être analyser.

Les deux dernières sections : Pour la construction de nos arbres de décision et le classifieur bayésien naïf et d'évaluer chaque modèle afin de comparer entre les performances des modèles et en déduire le meilleur entre ces modèles-là.

Logiciels et langages de programmation utilisés : Pour l'analyse graphique et descriptive nous avons utilisé *IBM SPSS Statistics 26* comme outil d'analyse, et le *langage de programmation python* a été utilisé pour la construction et évaluation des modèles.

Section 1 : Présentation et pré-traitement de la base de données et analyse descriptive

Pour des raisons de santé liées à la pandémie de COVID-19, et la situation des hôpitaux algériens durant cette pandémie, nous n'avons pas pu passer notre stage pratique ni collecter les données pour construire nos modèles, et pour cela nous avions donc trouver une alternative et avons élaboré un questionnaire pour faire une enquête en ligne afin de construire notre base de données qui nous servira à construire nos modèles.

1. Organisation et méthodologie de l'enquête

Objectifs de l'enquête

• Collecter des informations sur des femmes atteintes d'un cancer du sein et des femmes non atteintes, afin de voir les caractéristiques de chaque classe.

• La construction des modèles de classification en basant sur les facteurs de risque de cette maladie.

• La détermination du facteur de risque le plus discriminant de notre échantillon.

Type d'enquête et population cible

Pour des raisons qui ont été citées précédemment, nous avons opté pour une enquête en ligne, et pour la population cible nous avons essayé de cibler les femmes malades et non malades en cherchant en ligne (dans les réseaux sociaux : sur Facebook plus précisément) des groupes d'associations françaises du cancer du sein pour publier notre questionnaire et toucher les deux catégories de femme.

2. Présentation du questionnaire et des variables de la base de données

Questionnaire

Notre questionnaire (voir annexe 1) a été élaboré de manière qu'on puisse toucher tous les facteurs de risque qui ont été cité dans la deuxième section du

premier chapitre (voir pages 19-23), en formulant des questions de façon optimale, simple, et de manière courte, et facile à répondre.

Question	***Type***
Quel est votre âge ?	Question ouverte numérique.
Quel est votre poids ? (En kilogrammes)	Question ouverte numérique.
Quel est votre taille ? (En centimètres)	Question ouverte numérique.
Avez-vous un cancer du sein ?	Question fermée à choix unique.
Avez-vous un membre de votre famille atteint d'un cancer du sein ?	Question fermée à choix unique.
Si oui précisez (Mère, sœur, grand-mère, cousine, tente…)	Question à choix multiples.
Quelle est votre situation matrimoniale ?	Question fermée à choix unique.
Avez-vous des enfants ?	Question fermée à choix unique.
Si oui, combien ?	Question ouverte numérique.
A quel âge avez-vous votre premier enfant ?	Question ouverte numérique.
Avez-vous allaité vos enfants au sein ?	Question fermée à choix unique.
Si oui, quelle est la durée de votre allaitement ?	Question fermée à choix unique.
L'âge aux premières règles :	Question ouverte numérique.
À quel âge vous avez eu vos premières règles ?	Question ouverte numérique.
L'âge à la ménopause :	Question ouverte numérique.
Suivez-vous un traitement hormonal de la ménopause ?	Question fermée à choix unique.
Prenez-vous des pilules contraceptives ?	Question fermée à choix unique.
Pratiquez-vous une activité physique ?	Question fermée à choix unique.
Si oui, combien de fois par semaine ?	Question ouverte numérique.

Tableau (III-1) : Présentation des questions et leurs types.

Source : Elaboré par l'étudiant à l'aide de logiciel Excel2019.

Les variables initiales de la base de données

L'ensembles des questions de notre questionnaire va nous donner 19 variables qui se présentent comme suit :

Variable	*Type*
Id	Identifiant
Age	Numérique continue
Poids	Numérique continue
Taille	Numérique continue
Cancer_duSein	Booléenne (Variable cible)
Antécédents_familiaux	Booléenne
LeMembre_deFamille_atteint	Numérique discrète
Situation_Matrimoniale	Variable nominal
Parité	Booléenne
Nombre_D'enfants	Numérique discrète
Age_àLaPremNaiss	Numérique continue
Allaitement_AuSein	Booléenne
Durée_Allait	Variable ordinal
La_Ménarche	Numérique continue
Ageà_Laménopause	Numérique continue
THM	Booléenne
Pillules_contraceptives	Booléenne
Pratique_sport	Booléenne
Nombre_dePratSport	Numérique discrète

Tableau (III-2) : Présentation des variables initiales de la base de données.

Source : Elaboré par l'étudiant à l'aide de logiciel Excel2019.

Remarque : Ces variables-là sont des variables initiales de notre base de données, par la suite elles peuvent être modifier, transformer ou bien supprimer selon les besoins de notre prétraitement de la base de données.

3. Pré-traitement de la base de données

Cette étape est considérée comme une étape très importante et là nous allons utiliser un ensemble de techniques (cité dans le 2ème chapitre) pour convertir notre base de données brutes en un ensemble de données propre (voir Annexe 2).

Notre tableau de données brutes contient 247 instances (individus) ou répondants et 19 colonnes, dont une sert pour identifier chaque répondant (qui sera supprimer par la suite car dans le logiciel chaque répondant est numéroté automatiquement), et une deuxième colonne qui concerne notre variable d'intérêt (cible), et les autres

17 colonnes représentent les variables explicatives (attributs) de notre variables cible.

Et pour rendre notre tableau de réponses exploitables, plusieurs modifications ont été faites :

• La 1ère colonne qui concerne l'identification de chaque femme a été supprimer, car dans les logiciels seront numérotée automatiquement, et cette colonne ne donne pas d'informations importantes.

• Les deux variables Poids et taille ont été regroupées dans une seule variable (IMC) qui détermine la corpulence de chaque femme en basant sur les deux variables poids (en kilogrammes) et Taille (en mètres) et l'IMC est calculé comme suit :

$$IMC = \frac{Poids(kg)}{Taille^2(m)}$$

• Les réponses des deux questions (Avez-vous un membre de votre famille atteint d'un cancer du sein ? et Si oui précisez …) ont été fusionner dans une nouvelle colonne ou variable sous le nom (Nmbre_deMemFam_Malade) qui calcule le nombre de membres de famille malades pour chaque femme.

• Une nouvelle variable a été ajouté (Ménopausée) qui décrit si la femme est ménopausée (Oui) ou pas encore (Non), et cette nouvelle variable a été calculer à partir d'une fonction conditionnelle de la variable (âge à la ménopause).

• Les variables numériques continues (Age à la première naissance, âge à la ménopause) ont été transformer en variables ordinales afin de traiter les variables manquantes.

La liste finale des variables explicatives de notre tableau de données est la suivante :

Variable	*Description*
Cancer_DuSein	Si la femme est malade ou non (Variable cible)
Age	Variable qui mesure l'Age
IMC	Variable qui mesure la corpulence de chaque femme
Antecedent_Fami	Les antécédents familiaux liées au cancer du sein
Nmbre_deMemFam_Malades	Le nombre de membres de famille atteint du cancer du sein
Situation_Matrimoniale	La situation matrimoniale
Parite	La parité : si la femme a déjà accouché ou non
Nombre_d'enfants	Le nombre d'enfants
Age_aLaPrem_Naiss	L'âge de la femme à la première naissance
Allaitement_auSein	L'allaitement au sein
Duree_Allait	La durée d'allaitement
La_Menarche	L'âge des premières règles
Menopausee	Si la femme est ménopausée ou non
Age_Menopause	L'âge à la ménopause
THM	Traitement hormonal de la ménopause
Pills_contra	Si elle prend des pilules contraceptives
Activite_phys	Le pratique du sport
Nombre_dePrat	Le nombre de pratique du sport par semaine

Tableau (III-3) : La liste finale des variables explicatives.

Source : Elaboré par l'étudiant à l'aide de logiciel Excel2019.

Les modalités de chaque attribut sont détaillées dans l'annexe 3.

4. Analyse et visualisation des données

Aperçu et résumé statistique des caractéristiques des répondantes

Nous allons voir quelques caractéristiques de notre ensemble de 247 individues qui se résument comme suit :

Caractéristiques	Minimum	Maximum	Moyenne
Age	18	73	45
Taille (cm)	150	187	166,43
Poids (kg)	43	154	72,14
IMC (kg/m²)	15,7	61,67	26,07
Les antécédents familiaux (Nombre de proches malades)	1	3	1
Nombre d'enfants	1	7	2
Age aux premières règles	8	16	12
Age à la ménopause	35	60	50
Pratique de sport (nombre de fois /semaine)	1	7	3

Tableau (III-4) : Résumé statistique de quelques caractéristiques des répondantes.

Source : Elaboré par l'étudiant à l'aide de logiciel Excel2019.

• L'âge moyen de ces 247 femmes est de 45 ans, et varie entre 18 et 73 ans.

• le poids et la taille moyen(ne) sont de 72,14 kg et 166,43 cm, ce qui donnent un IMC moyen de 26,07 kg/m², qui est entre [25-30] et donc la moyenne des individus souffrent d'un surpoids.

• Presque 48% des répondantes ont au moins un membre de familles atteintes d'un cancer du sein (Arrière-grand-mère, grand-mère, mère, sœur, tante, cousine) avec une moyenne plus d'un membre malade.

• 84,6% de notre ensemble d'instances ont au moins accouchées une fois, et le nombre moyen des enfants pour chaque femme est de 2 enfants.

• La moyenne des âges aux premières règles est de 12 ans, et 36% des femmes sont ménopausées avec une moyenne de 50 ans à la ménopause, et parmi ces dernières que 11,2% qui suivent un traitement hormonal de la ménopause, et que 24,7% parmi l'ensemble de 247 femmes prennent des pilules contraceptives.

• Pour la pratique du sport presque 44,5% pratiquent une activé physique avec une moyenne de 3 fois par semaine.

Variable d'intérêt (cible)

Notre ensemble de 247 individus (instances) ou répondants est répartis selon la variable cible (Cancer_DuSein) en deux classes la première concerne les femmes non malades et la deuxième les femmes malades, et l'effectif de chaque classe est représenté comme suit :

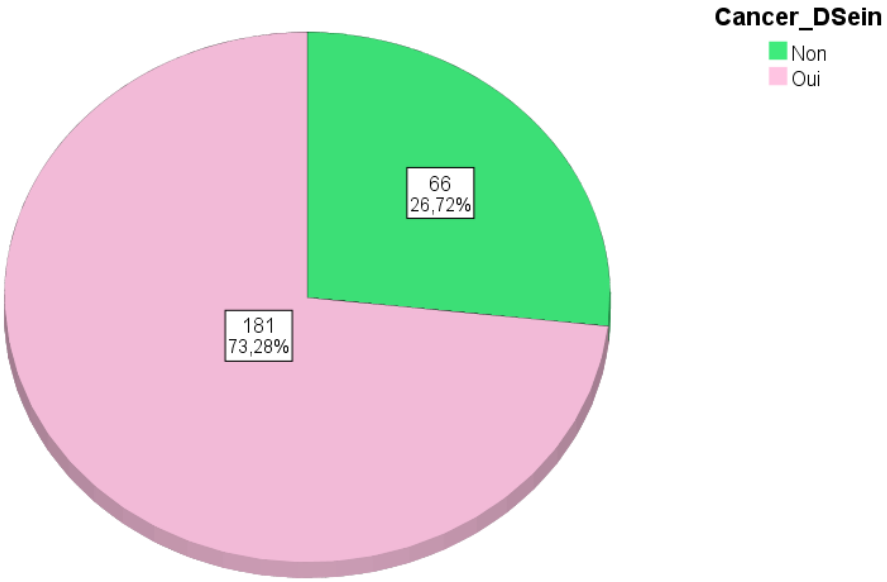

Figure (III-13) : Diagramme en secteur des effectifs des malades et non malades.

Source : Elaboré par l'étudiant à l'aide de logiciel IBM SPSS 26.

Comme le diagramme l'indique, la classe des femmes non malades représente 26,72% (soit 66 instances), et celle des femmes malades représente 73,28% (181 instances), ce qui nous donne deux proportions non équilibrées, ce qui va avoir une influence sur la précision et les performances de nos modèles. Et pour cela nous allons voir si nous utiliserons durant la construction des modèles des techniques pour équilibrer entre ces deux proportions ou bien non.

Variables explicatives

Là nous allons faire des analyses graphiques et descriptives sur l'ensemble des variables explicatives afin de voir comment chaque classe de notre variable cible se diffère selon les différentes variables explicatives et cela nous permet de comparer entre les caractéristiques des deux classes (les malades et les non malades).

L'âge

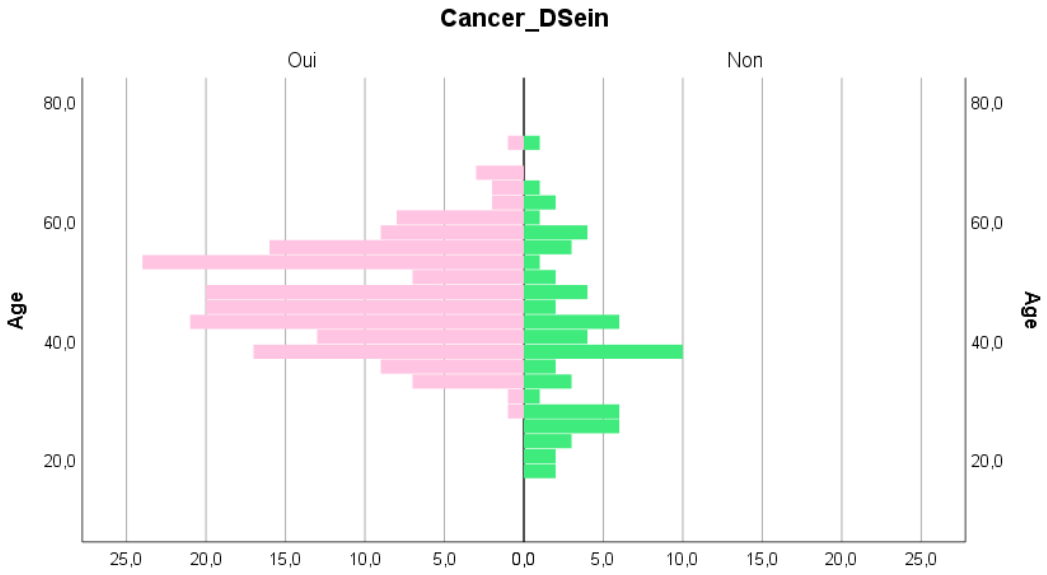

Figure (III-14) : Distribution des âges des malades et non malades.

Source : Elaboré par l'étudiant à l'aide de logiciel IBM SPSS 26.

Visuellement, on remarque que la moyenne d'âge des malades est plus élevée (autour de 50ans) que celle des non malades (environ 40ans) et pour qu'on puisse comparer entre les valeurs nous construirons un tableau descriptif :

Age (ans)	*Les non malades*	*Les malades*
Minimum	18	29
Maximum	73	72
Moyenne	39,8	47,5
Ecart-type	13	8,5

Tableau (III-5) : Tableau comparatif entre les âges des malades et non malades.

Source : Elaboré par l'étudiant à l'aide de logiciel Excel2019.

On a la moyenne des âges des non malades (39,8 ans) est inférieure à celle des malades (47,5 ans), ce qui confirmera ce que nous avons cité dans les facteurs du risque (dans le premier chapitre) que l'âge est le facteur de risque le plus important du cancer du sein, et l'incidence de cette maladie augmente avec l'âge.

IMC

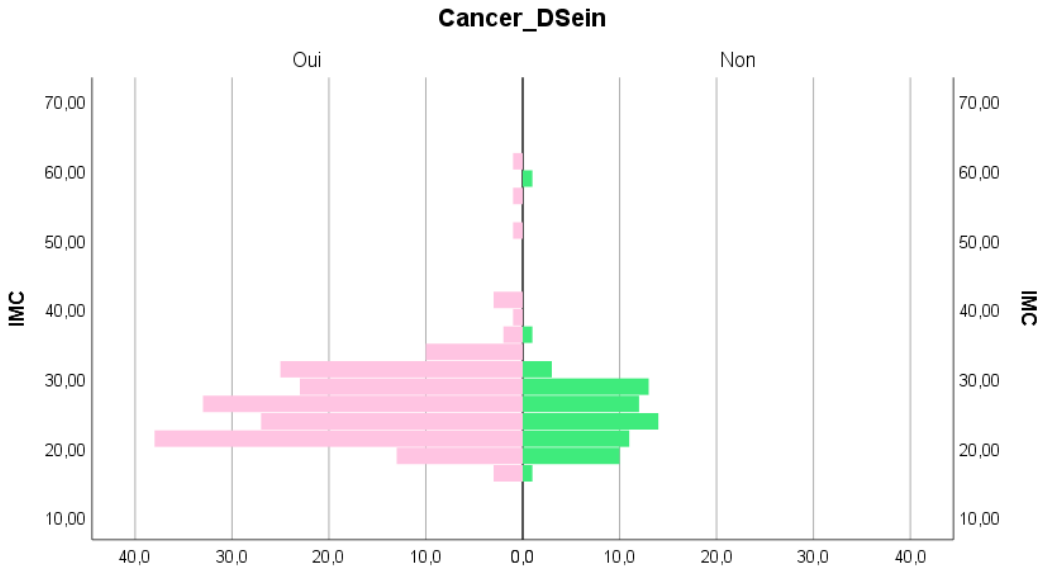

Figure (III-15) : Distribution des IMC des deux classes.

Source : Elaboré par l'étudiant à l'aide de logiciel IBM SPSS 26.

Graphiquement, les deux IMC des deux classes sont presque distribués de la même manière et ils sont autour de la même moyenne entre 20 et 30 kg/m² autour de 25kg/m², et pour différentier entre les deux classes nous allons faire une analyse descriptive :

IMC (kg/m²)	*Les non malades*	*Les malades*
Minimum	17 ,37	15,70
Maximum	55,10	61,67
Moyenne	24,95	26,48
Ecart-type	5,44	6,37

Tableau (III-6) : Tableau comparatif entre l'IMC des deux classes.

Source : Elaboré par l'étudiant à l'aide de logiciel Excel2019.

La classe des femmes malade est considérée comme une classe qui souffre d'un surpoids (IMC entre 25 et 30), ce qui augmentera le risque à développer certaines maladies (Diabète, cardiovasculaires, cancers). Et pour la classe des non malades (pour un IMC = 24,95 kg/m²) est considérée comme une classe normale avec un poids « santé », ce qui peut réduire le risque de développer certaines maladies.

Et cela confirmera ce que nous l'avons cité dans les facteurs de risque (dans le premier chapitre), « l'obésité est associée à un profil hormonal soupçonné de favoriser le développement du cancer du sein ».

Age à la première naissance

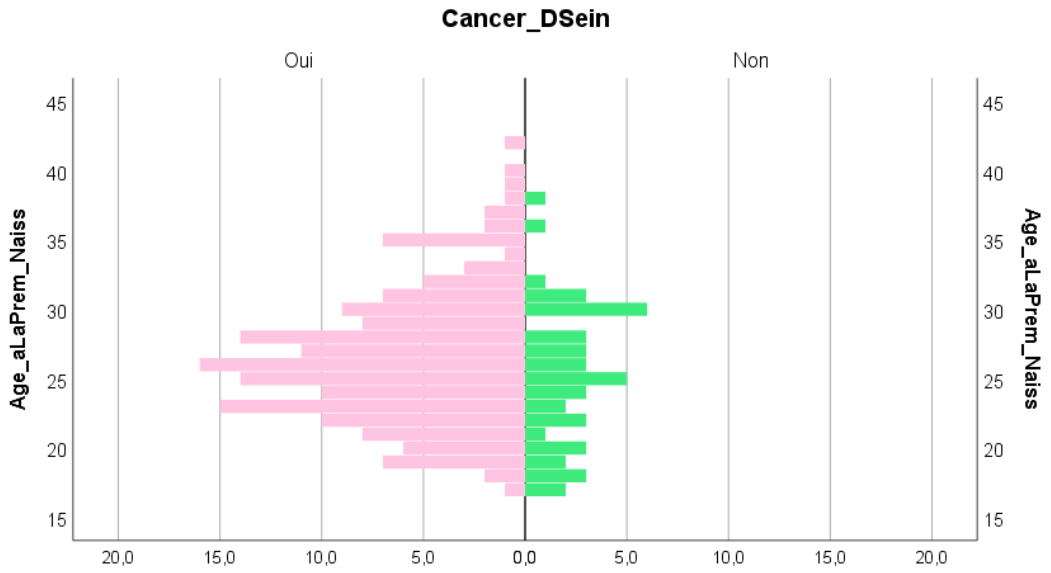

Figure (III-16) : La répartition des âges à la première naissance des deux classes.

Source : Elaboré par l'étudiant à l'aide de logiciel IBM SPSS 26.

Age à la première naissance (ans)	*Les non malades*	*Les malades*
Minimum	17	17
Maximum	38	42
Moyenne	25	27
Ecart-type	5	5

Tableau (III-7) : Tableau comparatif des âges à la première naissance de chaque classe.

Source : Elaboré par l'étudiant à l'aide de logiciel Excel2019.

D'après le tableau précédent on remarque que l'ensemble des non malades ont eu une première naissance à un âge moyen plus jeune que celle des malades, et la naissance à un âge avancé augmente le risque de développer un cancer du sein, et l'une des raisons serait que la grossesse change le tissu du sein de manière permanente, ce qui protège du cancer. Plus ce changement se produit tard après les premières règles de la femme, plus les cellules du sein ont du temps pour devenir cancéreuses.

La ménarche

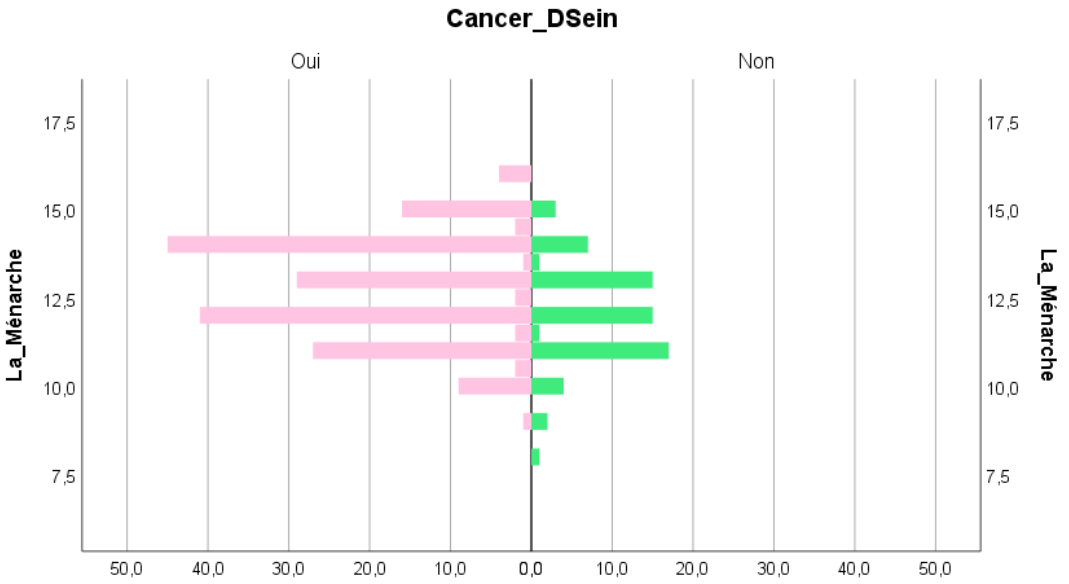

Figure (III-17) : distributions des âges aux premières règles des deux classes.

Source : Elaboré par l'étudiant à l'aide de logiciel IBM SPSS 26.

En visualisant le graphe on remarque que les âges aux premières règles des deux classes, sont autours 12,5 ans. Et pour faire une comparaison nous allons visualiser le tableau comparatif :

La ménarche (ans)	*Les non malades*	*Les malades*
Minimum	8	9
Maximum	15	16
Moyenne	12,1	12,8
Ecart-type	1,5	1,5

Tableau (III-8) : Tableau comparatif des âges aux premières règles des deux classes.

Source : Elaboré par l'étudiant à l'aide de logiciel Excel2019.

En étant un facteurs protecteur chaque année de retard dans l'installation des premières règles s'associe à une réduction de 5% du risque, et en analysant ce tableau on remarque que la moyenne des âges aux premières règles des malades 12,9 ans est inférieure à celle des non malades, malgré que cette différence ne soit pas énorme mais c'est peut-être dû l'inégalité entre les deux proportions des deux classes.

Age à la ménopause

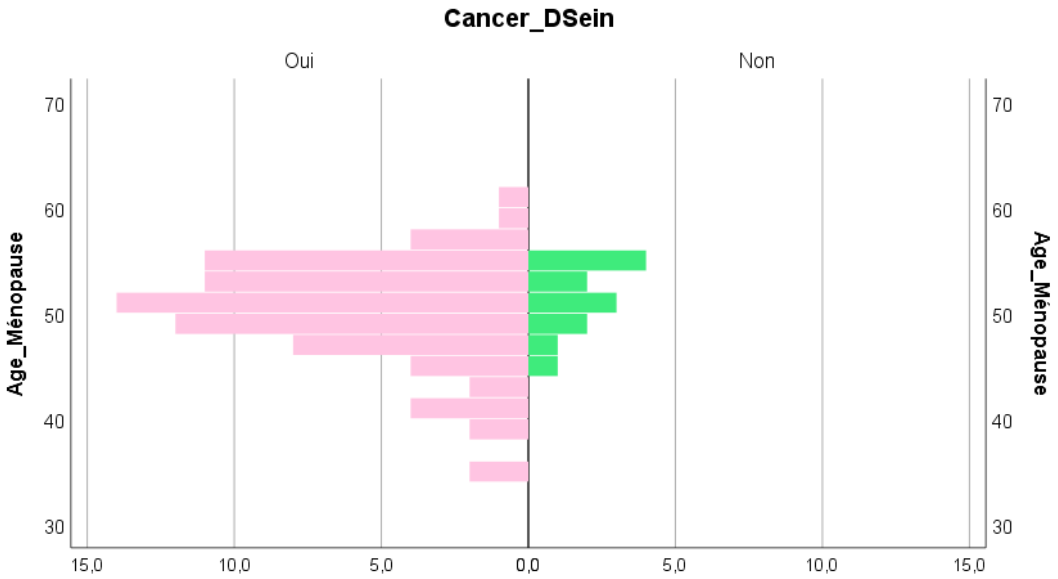

Figure (III-18) : Distribution des âges à la ménopause des deux classes.

Source : Elaboré par l'étudiant à l'aide de logiciel IBM SPSS 26.

Age à la ménopause	*Les non malades*	*Les malades*
Minimum	45	35
Maximum	55	60
Moyenne	51	49
Ecart-type	3	5

Tableau (III-9) : Tableau comparatif des âges à la ménopause des deux classes.

Source : Elaboré par l'étudiant à l'aide de logiciel Excel2019.

Comme on le regarde d'après le tableau et le graphe que la moyenne des âges à la ménopause des malades est bien inférieure que celle les non malades, et on sait que chaque retard année de retard dans l'installation de la ménopause s'associe à une augmentation de 3 à 4% du risque, mais cette comparaison est faussée car une bonne partie des femmes malade ont eu une ménopause forcée à cause du traitement (soit chimiothérapie ou hormonothérapie).

Et cette variable peut-être elle sera supprimée dans les étapes de la construction des modèles car elle donne des informations faussées.

La situation matrimoniale

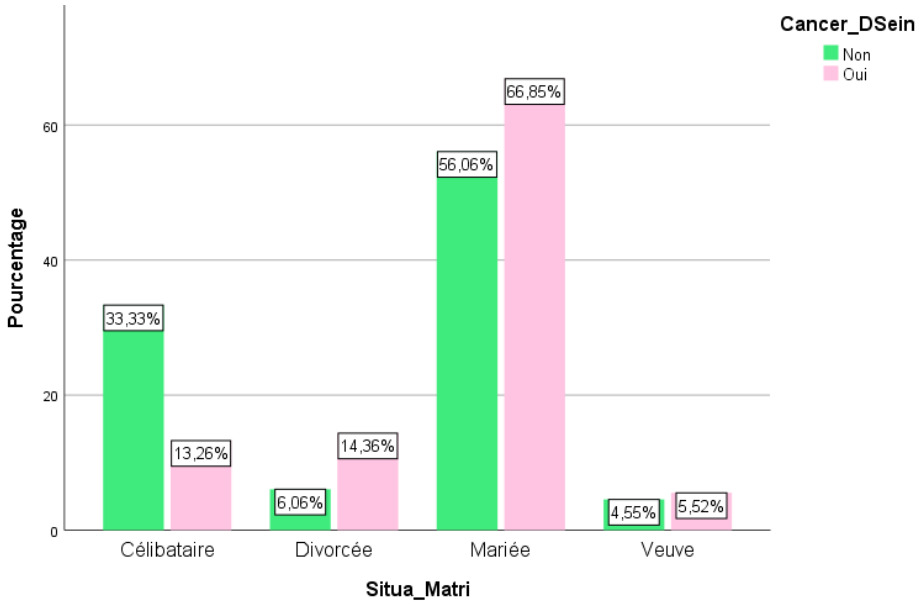

Figure (III-19) : La situation matrimoniale des deux classes.

Source : Elaboré par l'étudiant à l'aide de logiciel IMB SPSS 26.

Pour la situation matrimoniale, la majorité des répondantes des deux classes sont mariées (66,85% des malades et 56,06% des non malades), et pour les non malades un tiers sont célibataire et 13,26% des malades sont célibataires.

Antécédents familiaux et nombre de membres de famille malades

Nombre_Mem_de famille malades	*Les non malades*	*Les malades*
Minimum	0	0
Maximum	2	3
Moyenne	0	1
Ecart-type	1	1

Tableau (III-10) : Tableau comparatif des membres de famille malade des deux classes.

Source : Elaboré par l'étudiant à l'aide de logiciel Excel2019.

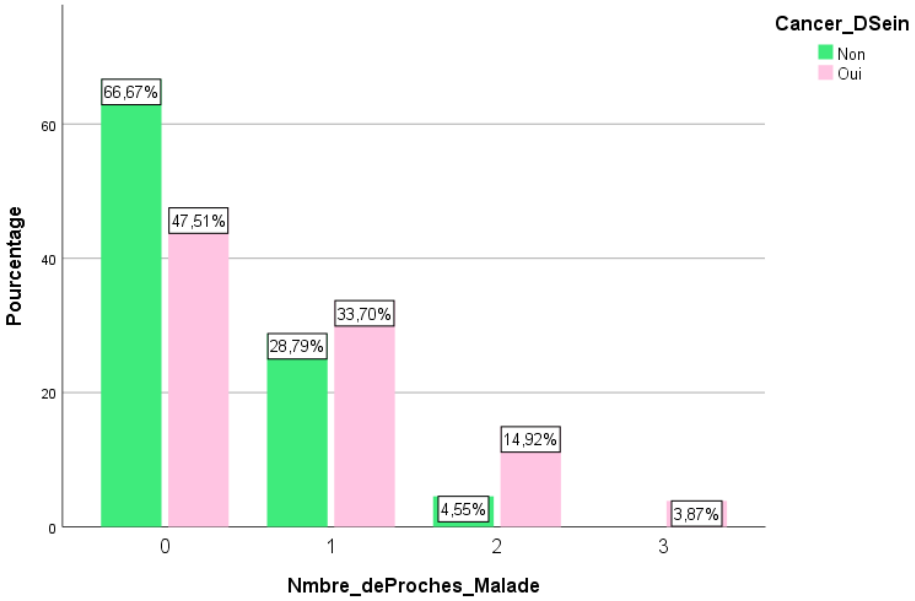

Figure (III-20) : Antécédents familiaux et nombre de proches malades des deux classes.

Source : Elaboré par l'étudiant à l'aide de logiciel IBM SPSS 26.

D'après le graphe et le tableau précédents on remarque que seulement 33,33% des non malades ont des antécédents familiaux, et plus de la moitié (52,49%) des femmes malades ont des antécédents familiaux avec une moyenne de deux proches malades, ce qui nous donnera une idée sur l'influence de la variable antécédents familiaux sur cette maladie.

Parité et nombre d'enfants

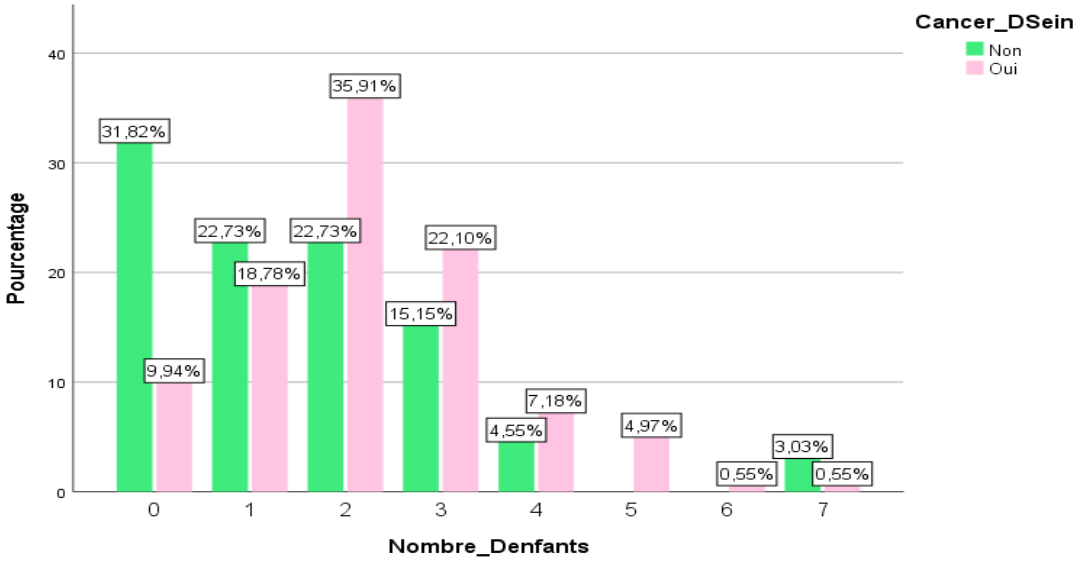

Figure (III-21) : Parité et nombre d'enfants des femmes malades et non malades.

Source : Elaboré par l'étudiant à l'aide de logiciel IBM SPSS 26.

Nombre d'enfants	Les non malades	Les malades
Minimum	0	0
Maximum	7	7
Moyenne	2	2
Ecart-type	2	1

Tableau (III-11) : Tableau comparatif de nombre d'enfants des deux classes.

Source : Elaboré par l'étudiant à l'aide de logiciel Excel2019.

Pour la parité et le nombre d'enfants des femmes, on remarque que les moyennes de nombre d'enfants des deux classes sont égales, et que 31,82% des femmes non malades sont nullipares, ce qui est liée peut être à la situation matrimoniale et l'âge des femmes non malades, et que 9,94% des femmes malades sont nullipares.

Durée et allaitement au sein

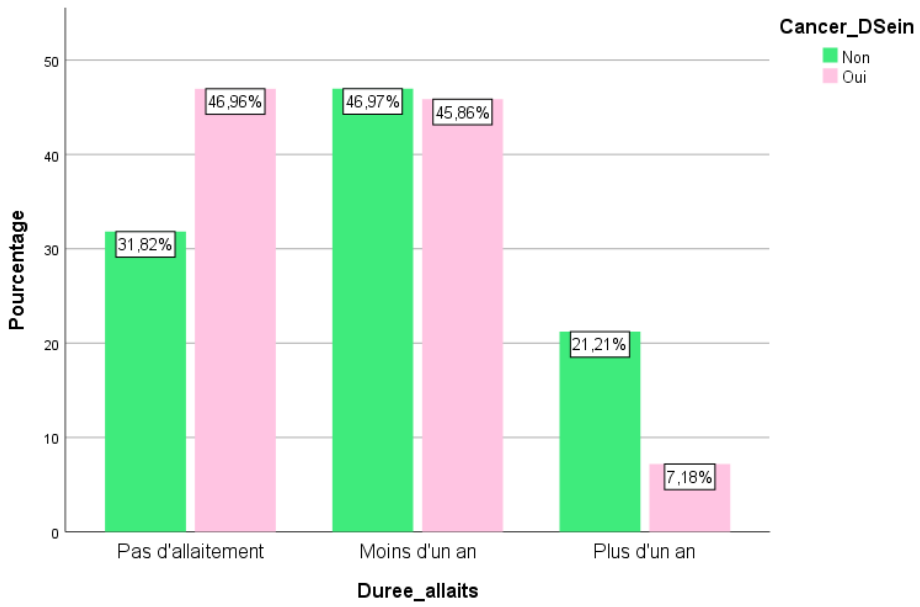

Figure (III-22) : Histogramme d'allaitement au sein et la durée pour les deux classes.

Source : Elaboré par l'étudiant à l'aide de logiciel IBM SPSS 26.

Pour l'allaitement au sein, on remarque que pour notre échantillon de 247 femmes, 46,96% des femmes malades n'allaitent pas au sein et seulement 31,82% des non malades qui n'allaitent pas, et aussi en visualisant le graphe on trouve que seulement 7,18% des femmes malades ont une durée d'allaitement plus d'un an, et 21,21% des femmes non malades allaitent pour une durée plus

d'un an. Alors on peut conclure pour notre échantillon que l'allaitement au sein est un facteur protecteur et un allaitement avec une durée plus d'un an diminuera le risque de développer un cancer du sein, ce qui confirmera ce que nous avons cités dans le premier chapitre.

Pilules contraceptives

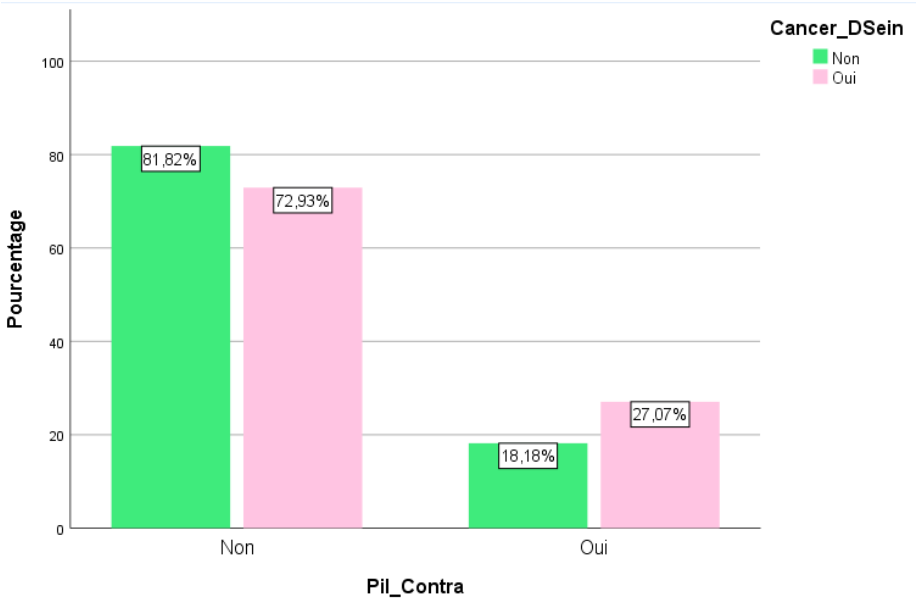

Figure (III-23) : Histogramme pour la variable pilules contraceptives.

Source : Elaboré par l'étudiant à l'aide de logiciel IBM SPSS 26.

Pour les pilules contraceptives, la majorité de notre échantillon ne prennent pas de pilules contraceptives, seulement 18,18% des non malades et 27,07% des malades qui prennent ces pilules.

Activité physique

D'après la représentation graphique ci-dessous : seulement 31,82% des non malades qui pratiquent une activité physique, et presque 50% des femmes malades pratiquent une activité physique.

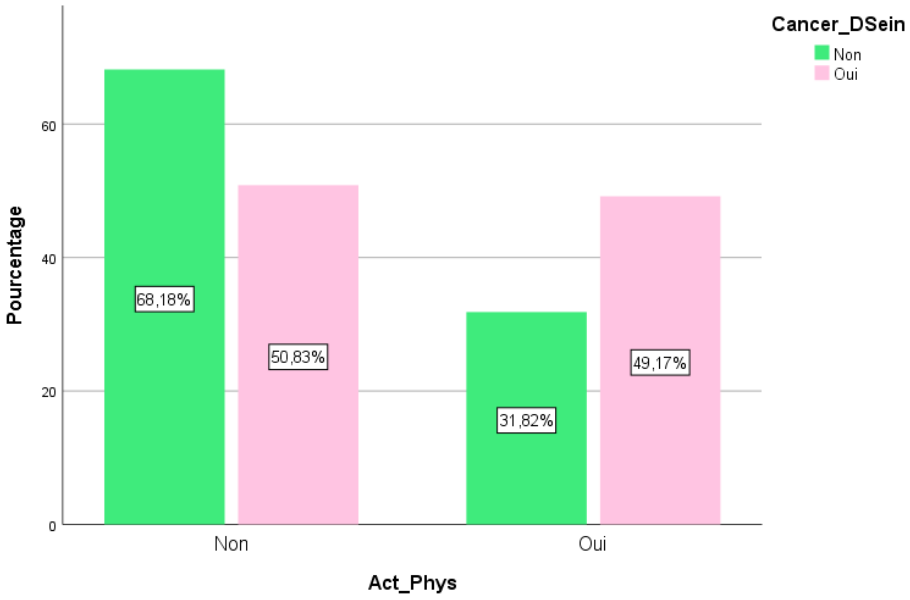

Figure (III-24) : Histogramme de la variable activité physique pour les deux classes.

Source : Elaboré par l'étudiant à l'aide de logiciel IBM SPSS 26.

Analyse des variables catégorielles

Dans cette partie, nous allons tester l'indépendance de nos variables catégorielles avec la variable cible (Cancer_DuSein), et pour faire cette analyse, nous allons regrouper nos variables et notre variable d'intérêt dans un tableau de contingence, qui est une représentation ou un croisement des deux modalités ou caractères en dénombrant l'effectif correspondant à la conjonction « $Modalités_1$ » et « $Modalités_2$ ».

Et pour vérifier cette indépendance nous testons les deux hypothèses suivantes par le test de Khi-deux :

H_0 : La variable d'intérêt Y est indépendante de la variable explicative X.

H_1 : La variable d'intérêt Y est dépendante de la variable explicative X.

Si la valeur calculée X^2_{Cal} est supérieure de la valeur tabulée X^2_{Tab}, on accepte l'hypothèse nulle (H_0).

Les résultats des différents tests d'indépendance avec la variable cible ont été fournis par le logiciel IBM SPSS 26, et ses résultats nous donnent la signification asymptotique bilatérale qui représente la probabilité que les deux

variables soient indépendantes. Si la signification asymptotique est inférieure à 5% (0,05), on conclut que les deux variables sont dépendantes et ses résultats sont résumés dans le tableau suivant (et comme les deux variables quantitatives l'âge à la ménopause et l'âge à la première naissance ont été transformer dans la construction des modèles en variables catégorielles, elles ont été inclues dans le tableau suivant) :

Variable	***Statistique de X^2***	***Degré de liberté***	***Signification asymptotique (bilatérale)***	***Décision***
Atecedent_Fami	7,1163	1	0,0076	Corrélées
Situation_Matrimoniale	14,1788	3	0,0027	Corrélées
Parite	15,3983	1	0,0001	Corrélées
Age_aLaPrem_Naiss	17,4767	2	0,0002	Corrélées
Allaitement_auSein	18,5887	3	0,0003	Corrélées
Duree_Allait	14,3692	3	0,0024	Corrélées
Menopausee	10,4272	1	0,0012	Corrélées
Age_Menopause	10,9911	2	0,0041	Corrélées
THM	7,5055	2	0,0062	Corrélées
Pills_contra	2,0553	1	0,1517	Non corrélées
Activite_phys	5,8960	1	0,0152	Corrélées

Tableau (III-12) : Présentation des résultats test de Khi-deux avec la variable cible.

Source : Elaboré par l'étudiant à l'aide de logiciel IBM SPSS 26.

Comme les résultats de ce tableau le montrent, notre variable cible (le cancer du sein) est dépendante avec toutes les variables catégorielles sauf avec les pilules contraceptives.

Section 2 : La classification par arbre de décision

Et dans cette section-là, que la construction de notre arbre de décision sera faite, à l'aide de Python 3.6 qui est un langage de programmation orienté objet, simple à utiliser et facile à apprendre. Avec ces différentes librairies qui faciliterons nos taches (Pré-traitement, codage, représentations graphiques, la construction et évaluation des modèles…).

Pour passer à l'étape de la construction de l'arbre de décision, un ensemble d'opérations est nécessaire à faire dans Python (Google Collab : environnement de travail pour l'écriture et l'exécution des programmes/codes) :

Étape 01 : l'importation des librairies nécessaires.

Étape 02 : l'importation des données (dans notre cas le fichier Excel qui contient les 247 réponses).

Étape 03 : Visualisation et recherche des variables manquantes et prétraitement des données (Ce qu'était fait dans la première section).

Étape 04 : La division des attributs en deux ensembles :

X : Les variables explicatives.

y : La variable cible.

Étape 05 : La sélection des variables explicatives (en basant sur la corrélation de Pearson et les variables qu'on trouve utile dans notre étude) et le codage des variables.

Étape 06 : La division des observations en deux groupes (70% pour construire nos modèles et 30% pour le test et la validation).

Apres que tous ces étapes sont faites on passera à l'application de notre algorithme :

Les paramètres de l'arbre de décision	***Spécification***	***Retenus***
Algorithme	*CART (Classification and regression trees)*	/
Type d'arbre	*Binaire*	/
Variable cible (dépendante)	Cancer_DuSein	/
Variables explicatives (indépendantes)	Age, IMC, Antecedent_Fami, Parite, Nombre_d'enfants, Age_aLaPrem_Naiss, Allaitement_auSein, Duree_Allait, La_Menarche, Pills_contra.	➢ Age, durée d'allaitement. (Modèle 1) ➢ Age, allaitement au sein et IMC (Modèle 2)
Critère de choix de la variable de segmentation	Le meilleur entre Gini et Entropy	Gini
L'élagage de l'arbre de décision	L'élagage avec les paramètres optimaux de l'arbre (GridSearchCV).	/
Profondeur maximale de l'arbre	Profondeur optimale	➢ 2 pour le modèle 1 ➢ 3 pour le 2eme modèle.
Validation	La technique du Train-test Split (70%, 30%)	172 individus pour la construction, Et 75 individus pour l'évaluation et validations des modèles.

Tableau (III-13) : Les paramètres optimaux pour la construction de l'arbre de décision.

Source : Elaboré par l'étudiant à l'aide de Excel2019.

1. La construction d'un arbre de décision préliminaire

L'arbre de décision

Premièrement, on commence par la construction d'un arbre de décision sans règle d'arrêt (sans élagage) afin de voir la précision de notre modèle préliminaire et d'avoir une idée sur le niveau d'élagage pour obtenir un arbre lisible et simple

et surtout de meilleure précision. (Et on a utilisé le Gini comme critère de séparation) :

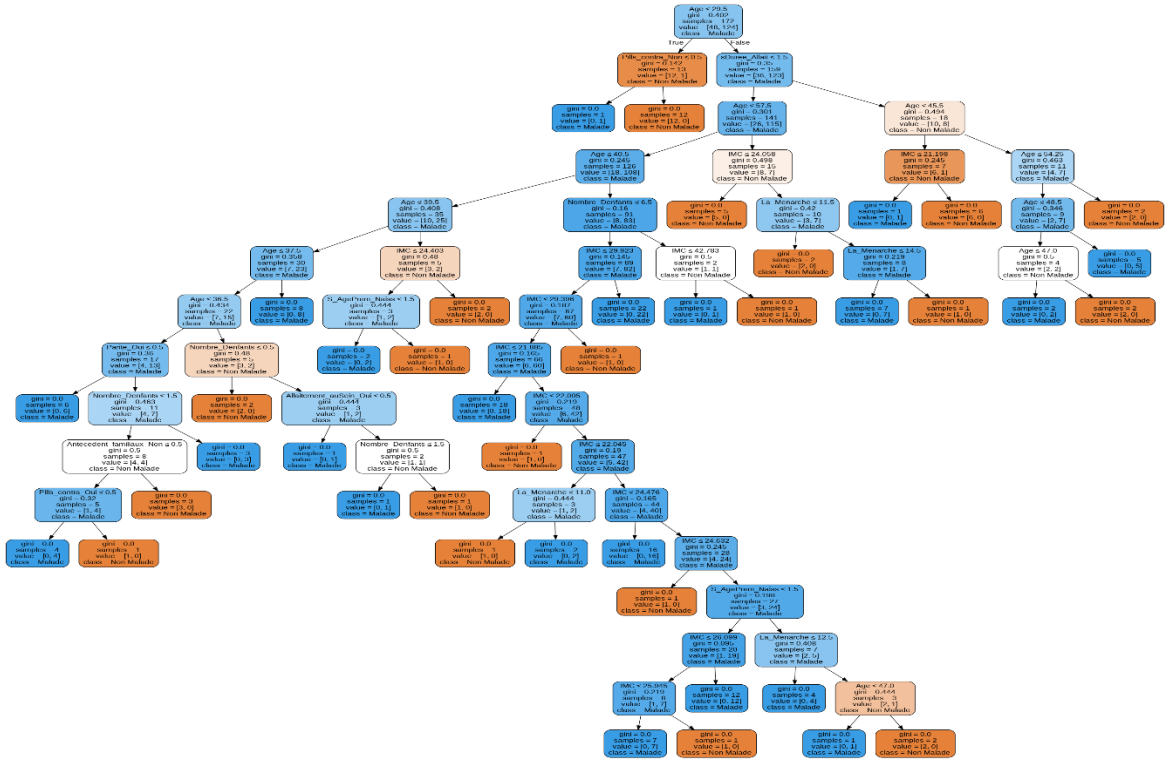

Figure (III-25) : Arbre de décision préliminaire (Non-élagué).

Source : Elaboré par l'étudiant à l'aide de Python.

Ce qui nous donne un arbre de décision non lisible et compliqué et long, car notre algorithme d'apprentissage cherche à construire une feuille pour chaque individu afin de classifier chaque individu sans erreur, ce qui va dégrader la performance de notre modèle et ce phénomène est appelée le Overfitting ou surapprentissage.

L'évaluation de l'arbre préliminaire

A l'aide de la matrice de confusion nous allons voir et mesurer les performances de cet arbre de décision en basant sur l'ensemble test (les 30% des individus).

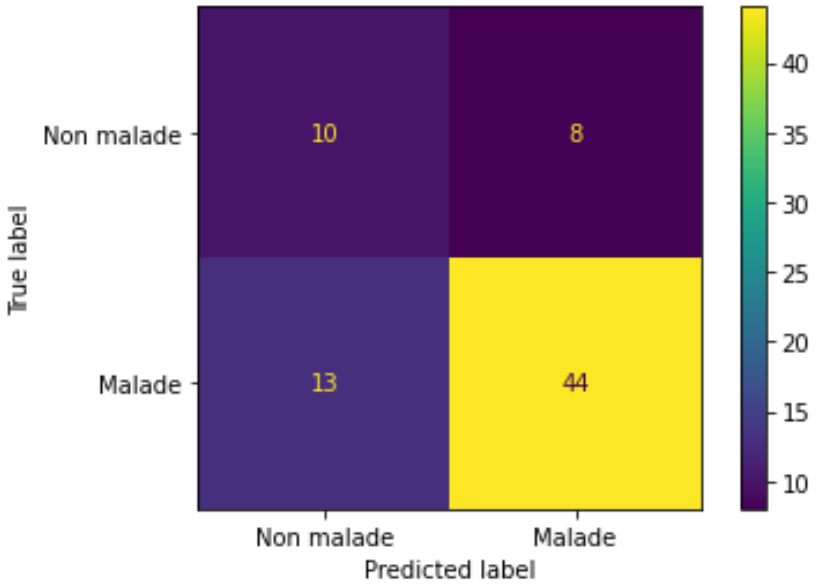

Figure (III-26) : Matrice de confusion de l'arbre de décision préliminaire.

Source : Elaboré par l'étudiant à l'aide de Python.

Et à partir de cette matrice on peut calculer quelques indicateurs de la performance :

Taux d'erreur=$\frac{FP+FN}{N+P}$ = 28%. (C'est la proportion des individus mal classées).

Sensibilité=$\frac{VP}{VP+FN}$ = 77,2%. (Notre modèle détecte 77,2% des malades).

Spécificité=$\frac{VN}{VN+FP}$ = 55,56%. (Il mesure la capacité du modèle à détecter les non malades).

Précision=$\frac{VP+VN}{N+P}$= 72%. La précision signifie que notre modèle arrive à prédire ou classer les individus avec une précision de 72%.

2. La construction de l'arbre de décision élagué

Comme nous l'avons cité dans le tableau (III-12), nous optons pour l'élagage à l'aide de la recherche des paramètres optimaux de l'arbre pour régler

le problème de Overfitting (surapprentissage), et nous allons faire ça à l'aide d'une technique (GridSearchCV) qui vérifie toutes les combinaisons des paramètres possible (Le critère de séparation, la profondeur de l'arbre, la taille minimale de séparation des nœuds enfants, et la taille minimale de chaque feuille), et qui nous donne la combinaison des paramètres optimale qui maximise la précision de notre arbre de décision.

L'arbre de décision (élagué)

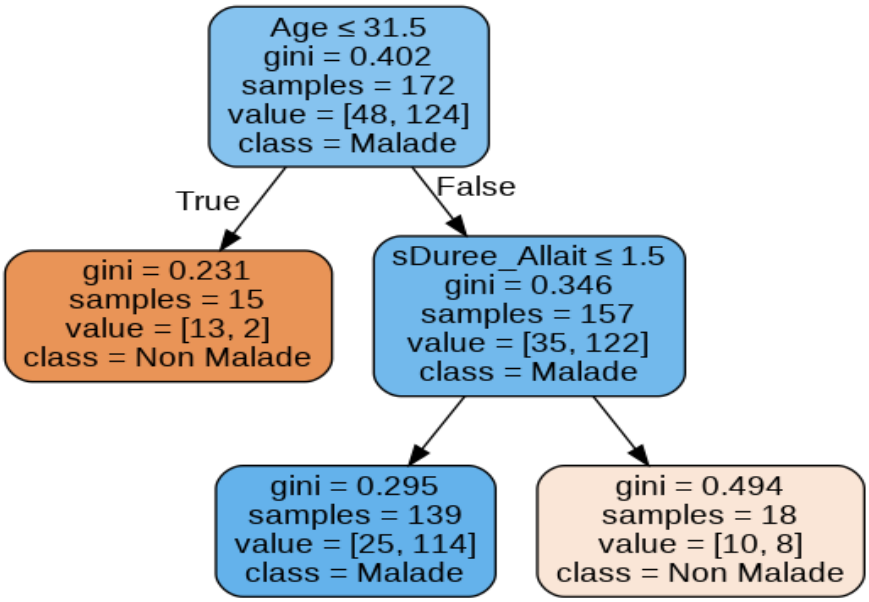

Figure (III-27) : Arbre de décision finale (Non équilibré).

Source : Elaboré par l'étudiant à l'aide de python.

Comme on regarde dans l'arbre précèdent que la variable explicative (Age) se situe dans le nœud racine, ce qui signifie que la variable Age est la meilleure variable de séparation (qui sépare la population en deux classes, les malades et non malades) selon le critère de Gini.

On peut déduire trois règles de décision sous la forme Si-Alors :

- Si l'âge est inférieur ou égale à 31,5 ans alors le risque de développer un cancer du sein est très bas (13,34%).

- Si l'âge est supérieur à 31,5 ans et la durée d'allaitement est moins d'un an ou la femme n'allait pas au sein alors le risque de développer un cancer du sein est très élevé (82,01%).

- Si l'âge est supérieur à 31,5 ans et la durée d'allaitement est plus d'un an alors le risque de développer un cancer du sein est bas (44,45%).

D'après l'arbre de décision (Figure III-17) et les 3 règles de décision précédentes, on peut déduire que l'âge est le facteur le plus important et le risque de développer un cancer du sein augmentera avec l'âge, et comme deuxième facteur le modèle a pris la durée d'allaitement (plus précisément une durée plus d'un an) qui est un facteur protecteur et la durée d'allaitement plus d'un an diminue le risque de développer un cancer du sein et augmentera si la durée est moins d'un an ou si les femmes n'allaitent pas.

L'évaluation de l'arbre de décision (élagué)

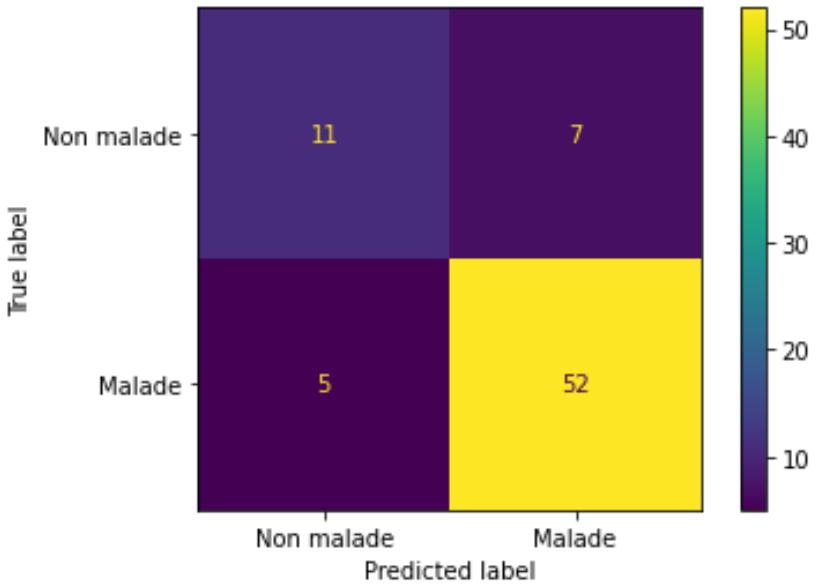

Figure (III-28) : Matrice de confusion de l'arbre finale.

Source : Elaboré par l'étudiant à l'aide de Python.

Les performances de notre arbre sont calculées à partir de la matrice de confusion et résumées dans le tableau suivant :

Indice de performance	Le taux
Taux d'erreur	16%
Sensibilité	92,23%
Spécificité	61,11%
Précision	84%

Tableau (III-14) : Tableau d'indice de performances de l'arbre de décision.

Source : Elaboré par l'étudiant à l'aide de Excel2019.

La précision de notre modèle élagué (84%) est beaucoup meilleure que celle de l'arbre préliminaire (72%) et la capacité de notre modèle de détecter les malades a augmenté (77,2% contre 92,23%), et le taux d'erreur (ou la mauvaise classification) de notre modèle a diminué. Par contre la sensibilité est faible 61,11% ce qui signifie que notre modèle n'arrive pas à détecter les non malades, et cela c'est dû à l'inégalité entre les deux proportions des malades et non malades, ce qui a influencé les performances de notre modèle, il détecte mieux les malades que les non malades.

3. La construction d'un arbre de décision équilibré

Pour fixer ce problème nous avons essayé d'équilibrer entre les proportions des malades et non malades pour comparer entre les performances des deux modèles (avant équilibrage et après).

En utilisant une technique d'Under-sampling (sous échantillonnage) qui consiste à tirer un échantillon aléatoire de la classe majoritaire (les malades) d'une même taille que la classe minoritaire (les non malades) afin de construire notre modèle sur deux proportions égales.

L'arbre de décision (équilibré)

Apres que le tirage a été fait et les deux proportions sont égales, nous cherchons les paramètres optimaux pour construire notre arbre de décision à l'aide de la méthode GridSearchCV :

Critère : Gini ; la profondeur maximale : 3 ; la taille minimale des feuilles : 15.

Et l'application de ces paramètres donnera l'arbre de décision suivant :

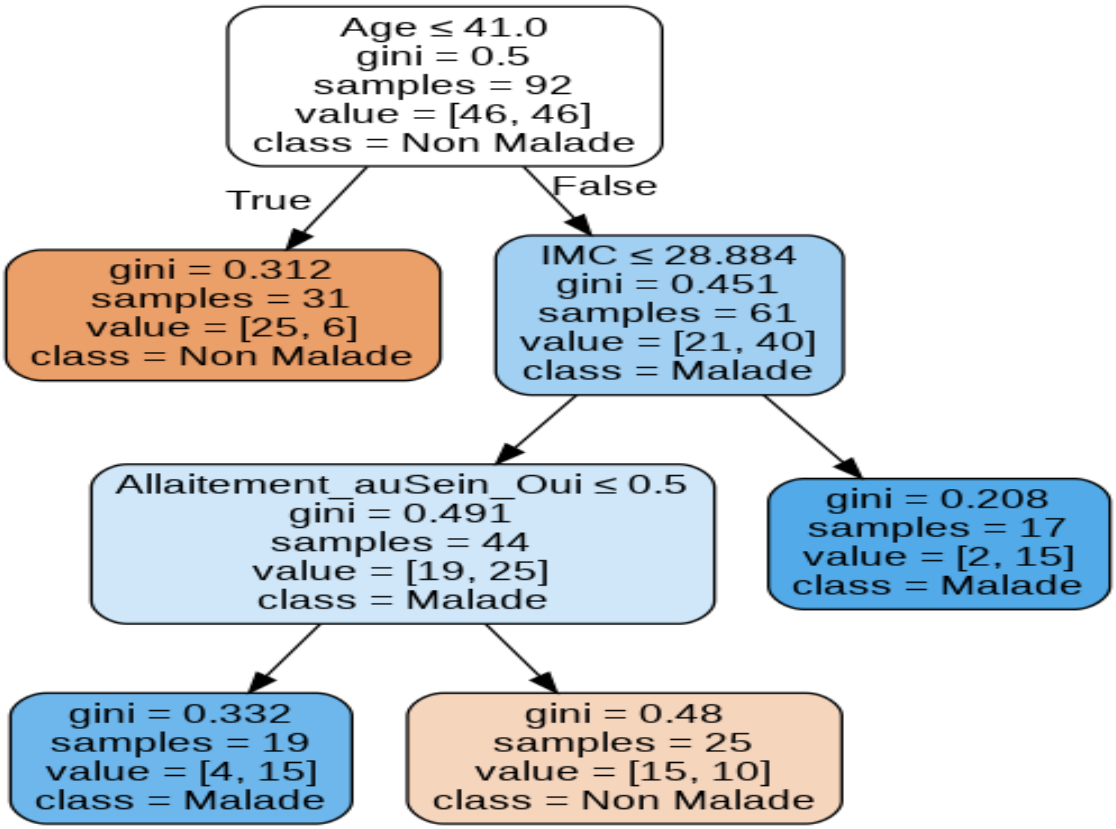

Figure (III-29) : L'arbre de décision finale (équilibré).

Source : Elaboré par l'étudiant à l'aide de Python.

Comme l'arbre précédent l'âge est le facteur le plus important et qui partage le mieux les malades et les non malade, et qui minimise le critère de Gini, mais cette fois-ci on remarque la présence de deux autres facteurs l'IMC et l'allaitement au sein et que les proportions sont équilibrées, et à partir de cet arbre nous allons construire quatre règles de décision :

- Si l'âge est inférieur ou égale à 41 ans alors le risque de développer un cancer du sein est 19,35%.

- Si l'âge est supérieur à 41 ans et l'IMC est supérieure à 28,88 kg/m², alors le risque de développer un cancer du sein est très élevé 88,23%.

- Si l'âge > 41 ans et IMC ≤ 28,88 kg/m² et l'allaitement au sein égale à oui alors le risque de développer un cancer du sein est bas et égale à 40%.

- Si l'âge > 41 ans et l'IMC ≤ 28,88 kg/m² et la femme n'allaite pas au sein alors le risque est élevé 78,94%.

D'après l'arbre de décision et ses quatre règles de décision, on peut conclure que :

- L'âge en premier lieu est le facteur du risque le plus important, le risque diminuera avec la diminution de l'âge et vice versa.
- L'indice de la masse corporelle (IMC) ou l'obésité joue un rôle très important, et le risque de développer un cancer du sein est plus important lorsque l'IMC est supérieur à 29 kg/m², 88% quand l'IMC est supérieur contre 56% lorsque l'IMC ≤ 29 kg/m².
- L'allaitement au sein est un facteur protecteur, il diminuera le risque de développer le cancer du sein de presque 40%.

L'évaluation de l'arbre de décision (équilibré)

Tous d'abord nous allons construire la matrice de confusion de notre modèle :

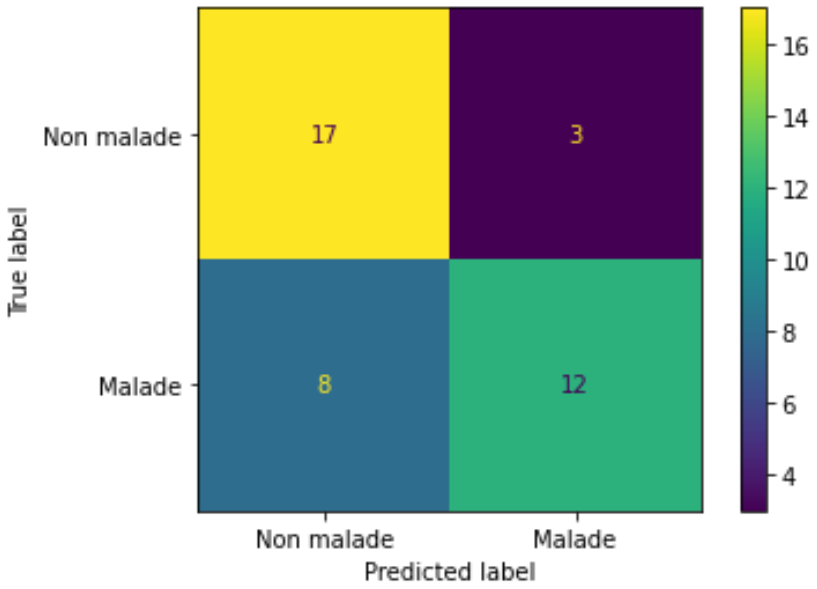

Figure (III-30) : Matrice de confusion de l'arbre équilibré.

Source : Elaboré par l'étudiant à l'aide de Python.

A partir de cette matrice de confusion nous allons évaluer notre arbre de décision :

Indice de performance	*Le taux*
Taux d'erreur	***27,5%***
Sensibilité	***60%***
Spécificité	***85%***
Précision	***72,5%***

Tableau (III-15) : Tableau de performance de l'arbre de décision équilibré.

Source : Elaboré par l'étudiant à l'aide de Python.

Comme le tableau l'indique, la précision de notre modèle est égale à 72,5% ce qui est entre les précisions des deux autres modèles, et notre modèle arrive à détecter les non malades mieux que les malades (un taux de 85% contre 60%), et la proportion des personnes mal classé est égale à 27,5, ce qui est un petit peu élevé.

Section 3 : La classification bayésienne naïve

Après que nous avons construits et déterminer avec les arbres de décision les facteurs de risque les plus discriminants de notre échantillon, l'idée-là est de construire un modèle de classification basés sur les calculs de probabilités qui nous permet de classifier les femmes en deux classes les malades et les non malade selon un ensemble de caractéristique (les facteurs de risque) et pour cela nous allons construire un classifieur bayésien naïf.

1. La construction du modèle

Pour la construction du classifieur bayésien naïf nous allons suivre les étapes suivantes :

- La division des attributs en X (pour les variables explicatives) et y (pour la variable cible).

- Le partage de l'ensemble des individus en deux groupes 70% pour la construction du modèle et 30% pour l'évaluation et la validation.

- La standardisation des variables quantitatives (par le StandardScaler), qui consiste à normaliser l'ensemble des variables explicatives comme suit :

$$z = \frac{x - \mu}{\sigma}$$

- Pour la présence des variables quantitatives nous optons pour un classifieur bayésien naïf gaussien.

Les variables explicatives qui ont été prisent en compte pour la construction de notre classifieur sont les suivantes :

Variables quantitatives	*Variables qualitatives*
Age, IMC, Nmbre_deProches_Malade, Nombre_Denfants, La_Menarche.	Antecedent_familiaux, Parite, Age_aLaPrem_Naiss, Allaitement_auSein, Duree_Allait, Pills_contra.

Tableau (III-16) : Tableau des variables de construction du modèle bayésien naïf.

Source : Elaboré par l'étudiant à l'aide de Excel2019.

Donc le principe de notre classifieur bayésien naïf est de calculer les probabilités conditionnelles d'appartenance d'une femme ayant des caractéristiques X (dans notre cas les facteurs de risque cités dans le tableau précèdent) à une classe (Soit malade ou bien non malade), en tenant compte l'hypothèse d'indépendance des variables explicatives X avec la variables d'intérêt y, puis on prend la classe qui maximise cette probabilité et attribuer notre individu à cette classe (les malades ou non malades).

Et pour mieux comprendre comment la classification bayésienne naïve gaussienne fonctionne, nous allons prendre les caractéristiques des deux femmes (sélectionnées aléatoirement à partir de l'échantillon test) et nous allons calculer les probabilités d'appartenir à chaque classe et de comparer entre la classe prédite et la classe réelle :

Femme 01 : âge= 39 ; IMC=24, 69 ; Antécédents familiaux= Oui ; Parité= Oui ; Nombre d'enfants= 02 ; âge à la première naissance >=30 ; allaitement au sein= Oui ; durée allaitement= Moins d'un an ; La ménarche= 11 ; pilules contraceptives= Non.

Femme 02 : âge= 51 ; IMC=21,79 ; Antécédents familiaux= Oui ; Parité= Oui ; Nombre d'enfants= 02 ; âge à la première naissance <30 ; allaitement au sein= Oui ; durée allaitement= Moins d'un an ; La ménarche= 13 ; pilules contraceptives= Oui.

Calcule des probabilités

Pour chaque femme nous allons calculer la probabilité d'appartenance dans chaque classe en suivant la formule suivante :

$$P(C_i/X = x_1, x_2, \ldots . x_p) = \prod_{j=1}^{p} P(x_j/C_i) * P(C_i)$$

C_1 : la classe des malades, C_2 : la classe des non malades.

X_i : l'ensemble des caractéristiques de l'individu *i*.

Pour la 1ère femme :

P (Malade | X_1= *{âge=39, IMC=24,69……, pilules contraceptives=Non}*)

$= \prod_{j=1}^{11} P(xj \mid Malade) * P(Malade)$ = 0,0004.

P (Non malade | X_1= *{âge=39, IMC=24,69……, pilules contraceptives=Non}*) = $\prod_{j=1}^{11} P(xj \mid Non\ malade) * P(Non\ malade)$

= 0,00000784.

On a : P (Malade | X_1) = 0,0004> P (Non malade | X_1) = 0,00000784

Alors : cette femme est classée selon notre classifieur bayésien naïve gaussien comme étant Malade.

Pour la 2ème femme :

P (Malade | X_2) = 0,000306 > P (Non malade | X_2) = 0,0000134

Alors : cette femme sera classée comme étant malade.

Et la question ici comment ces probabilités sont calculées ?

Les probabilités sont calculées à partir des distributions conditionnelles des variables explicatives avec la variable cible (calculées à partir de l'échantillon d'apprentissage) :

Les variables qualitatives

La probabilité des classes

Classe	*Effectif*	*Probabilités*
Malade	124	0,720930233
Non malade	48	0,279069767
Total	172	1

Tableau (III-17) : Représentation des probabilités des classes.

Source : Elaboré par l'étudiant à l'aide de Excel2019.

Comme le tableau l'indique, notre échantillon d'apprentissage est réparti d'une façon non équiprobable, on a presque trois quarts des femmes qui sont atteintes d'un cancer du sein.

A partir de ce tableau on peut calculer les probabilités des deux classes : P (Non malade) = 0,28 Et P (Malade) = 0,72.

Antécédents familiaux

Dans le tableau suivant nous allons voir la distribution de la variable antécédents familiaux selon la variable cible :

Classe/Antécédent familiaux	*Non*	*Oui*	*Total*	*P (Non | C)*	*P (Oui | C)*
Non malade	35	13	48	0,73	0,27
Malade	68	56	124	0,55	0,45
Total	103	69	172	0,60	0,40

Tableau (III-18) : Représentation de la probabilité conditionnelle P (Antécédent familiaux \| Classe).

Source : Elaboré par l'étudiant à l'aide de Excel2019.

Ce tableau nous permet de calculer les probabilités conditionnelles P (Antécédent familiaux | Classe), et d'après ce tableau on remarque que 40% de notre échantillon d'apprentissage ont des antécédents familiaux, et que parmi les 124 malades 45% ont des antécédents familiaux.

Parité

Ce tableau nous présente les probabilités conditionnelles pour les modalités de la variable parité P ($X=x_j$ | $Y=C_i$) :

Classe/Parité	*Non*	*Oui*	*Total*	*P (Non | C)*	*P (Oui | C)*
Non malade	12	36	48	0,25	0,75
Malade	11	113	124	0,09	0,91
Total	23	149	172	0,13	0,87

Tableau (III-19) : Représentation de la probabilité conditionnelle P (Parité | Classe).

Source : Elaboré par l'étudiant à l'aide de Excel2019.

Selon ce tableau, on remarque que 87% de notre échantillon d'apprentissage ont au moins accouchées une fois, et 91% parmi les malades et 75% parmi les non malades.

Age à la première naissance

Ce tableau nous présente les probabilités conditionnelles pour les modalités de la variable âge à la première naissance P ($X=x_j$ | $Y=C_i$) :

Classe/Age à la première naissance	*<30*	*>=30*	*Pas encore*	*Total*	*P (<30 | C)*	*P (>=30 | C)*	*P (Pas encore | C)*
Non malade	26	9	13	48	0,54	0,19	0,27
Malade	87	26	11	124	0,70	0,21	0,09
Total	113	35	24	172	0,66	0,20	0,14

Tableau (III-20) : Représentation de la probabilité conditionnelle P (Age à la première naissance | Classe).

Source : Elaboré par l'étudiant à l'aide de Excel2019.

Selon ce tableau, on observe que les femmes (de notre échantillon d'apprentissage) qui ont une naissance à un âge<30 ont une forte probabilité conditionnelle d'être classées dans la classe des malades.

Allaitement au sein

A partir de tableau suivant que nous allons calculer les probabilités conditionnelles de la variable allaitement au sein avec nos deux classes :

Classe/Allaitement	Non	Oui	Pas d'enfants	Total	P (Non | C)	P (Oui | C)	P (Pas d'enfants | C)
Non malade	9	27	12	48	0,19	0,56	0,25
Malade	47	67	10	124	0,38	0,54	0,08
Total	56	94	22	172	0,32	0,55	0,13

Tableau (III-21) : Représentation de la probabilité conditionnelle P (Allaitement au sein \| Classe).

Source : Elaboré par l'étudiant à l'aide de Excel2019.

Pour les résultats obtenus dans ce tableau, on observe que parmi les femmes malade 38% n'allaitent pas au sein et pour les femmes non malades on a que 19% qui n'allaitent pas.

Durée allaitement

Ce tableau nous présente les probabilités conditionnelles pour les modalités de la variable durée d'allaitement $P(X=x_j \mid Y=C_i)$:

Classe/Durée allaitement	Moins d'un an	Plus d'un an	Pas d'allaitement	Total	P (Moins d'un an| C)	P (Plus d'un an| C)	P (Pas d'allaitement| C)
Non malade	17	10	21	48	0,35	0,21	0,44
Malade	60	8	56	124	0,48	0,07	0,45
Total	77	18	77	172	0,45	0,10	0,45

Tableau (III-22) : Représentation de la probabilité conditionnelle P (Durée allaitement \| Classe).

Source : Elaboré par l'étudiant à l'aide de Excel2019.

Comme on observe dans le Tableau (III-21) que l'allaitement au sein avec une période plus d'un an diminue le risque de développer un cancer du sein, pour 124 personne malade que 7% parmi eux qui ont une période d'allaitement plus d'un an.

Pilules contraceptives

Ce tableau nous présente les probabilités conditionnelles pour les modalités de la variable pilules contraceptives $P(X=x_j \mid Y=C_i)$:

Classe/Pilules contraceptives	*Non*	*Oui*	*Total*	*P (Non | C)*	*P (Oui | C)*
Non malade	40	8	48	0,83	0,17
Malade	89	35	124	0,72	0,28
Total	129	43	172	0,75	0,25

Tableau (III-23) : Représentation de la probabilité conditionnelle P (Pilules contraceptives \| Classe).

Source : Elaboré par l'étudiant à l'aide de Excel2019.

D'après ce tableau, on remarque que seulement 25% de notre échantillon test qui prennent des pilules contraceptives, et parmi les 124 malades 28% prennent des pilules contraceptives.

Les variables quantitatives

Pour les variables quantitatives, les probabilités seront calculées à partir de la fonction de densité (Likelihood) en appliquant la fonction de densité de la loi normale suivante :

$$f(x) = \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{(x-m_i)^2}{2{\sigma_i}^2}}$$

Avec σ_i : l'écart type de la classe *i*, et m_i : la moyenne de la classe *i.*

Variables	Non malade (Moyenne)	Non malade (écart-type)	Malade (Moyenne)	Malade (écart-type)
Age	41,948	13,185	46,685	8,12
IMC	25,151	5,979	26,402	6,503
La ménarche	12,188	1,482	12,613	1,513
Nombre d'enfants	1,77	1,666	2,21	0,894

Tableau (III-24) : Caractéristiques des variables quantitatives des deux classes (échantillon d'apprentissage).

Source : Elaboré par l'étudiant à l'aide de Excel2019.

Et comme nous l'avons cité dans les étapes de la construction du classifieur bayésien naïf gaussien nous allons travailler avec des données centrées et réduites (Avec m=0, et σ=1).

Et là nous allons reprendre les mêmes caractéristiques des deux femmes pour recalculer les probabilités à partir des tableaux des probabilités conditionnelles :

Femme 01 : âge= 39 ; IMC=24, 69 ; Antécédents familiaux= Oui ; Parité= Oui ; Nombre d'enfants= 02 ; âge à la première naissance >=30 ; allaitement au sein= Oui ; durée allaitement= Moins d'un an ; La ménarche= 11 ; pilules contraceptives= Non.

P (Non malade | X_1= *{âge=39, IMC=24,69……., pilules contraceptives=Non}*) = $f_{Non\ malade}$ (Age=39) *$f_{Non\ malade}$ (IMC=24,69) *P (Antécédents familiaux=Oui | Non malade) *P (Parité=Oui | Non malade) *$f_{Non\ malade}$ (Nombre d'enfants=02) *P (Age à la première naisse>=30 | Non malade) *P (Allaitement au sein=Oui | Non malade) *P (Durée allaitement=Moins d'un an | Non malade) *$f_{Non\ malade}$ (La ménarche=11) *P (Pilules contraceptives= Non | Non malade) * P (Non malade)

=0,38*0,39*0,27*0,75*0,386*0,19*0,56*0,35*0,28*0,83*0,28*0,28

= 0,00000784.

P (Malade | X_1) = 0,24*0,38*0,45*0,91*0,38*0,21*0,54*0,48*0,22*0,72*0,72
= 0,0004.

On a : P (Malade|X_1) > P (Non malade|X_1), alors cette femme sera classée comme étant malade.

Femme 02 : âge= 51 ; IMC=21,79 ; Antécédents familiaux= Oui ; Parité= Oui ; Nombre d'enfants= 02 ; âge à la première naissance <30 ; allaitement au sein= Oui ; durée allaitement= Moins d'un an ; La ménarche= 13 ; pilules contraceptives= Oui.

P (Non malade | X2) =
0,31*0,33*0,27*0,75*0,39*0,54*0,56*0,35*0,33*0,17*0,28 = 0,0000134.

P (Malade | X2) =0,34*0,3*0,45*0,91*0,38*0,7*0,54*0,48*0,38*0,28

= 0,000306.

On a : P (Malade|X_1) > P (Non malade|X_1), alors cette femme sera classée comme étant malade.

2. L'évaluation du classifieur bayésien naïf

Après la construction de notre modèle nous allons évaluer ses performances de notre modèle sur l'ensemble de test (les 30%) à partir de la matrice de confusion suivante :

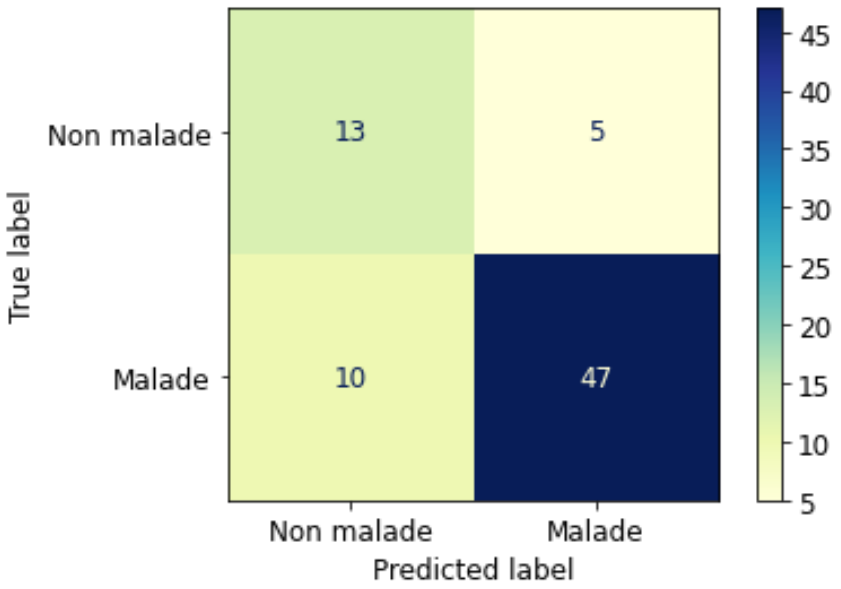

Figure (III-31) : Matrice de confusion (classifieur bayésien naïf gaussien).

Source : Elaboré par l'étudiant à l'aide de Python.

A partir de la matrice de confusion on remarque que le classifieur bayésien naïf gaussien donne de bons résultats avec une précision de 80% et un taux d'erreur de 20%, et le modèle arrive à détecter 82,45% des malades ce qui est bien, et il a bien classer 72,23% des non malades ce qui nous donnera un ensemble de performance acceptables, malgré que le classifieur bayésien naïve repose sur l'hypothèse d'indépendance des variables explicatives avec la variable d'intérêt, le modèle a de bonnes performances.

Conclusion

Dans ce dernier chapitre nous avons présenté notre ensemble d'individu et analyser ses différentes caractéristiques, puis la construction de nos modèles a été faite, et nous avons obtenus les résultats suivants (Pour notre échantillon de 247 femmes) :

- Pour le premier arbre de décision, on a vu que l'âge est le facteur de risque le plus importants et la durée d'allaitement au sein plus d'un an diminuera le risque de développer un cancer du sein.

- Et pour le deuxième arbre de décision, les facteurs les plus discriminants sont l'âge et l'indice de la masse corporelle, et l'allaitement au sein est considéré comme un facteur protecteur.

- Les résultats obtenus du classifieur bayésien naïf sont acceptables avec 80% de précision, et il arrive à détecter les malades avec un taux presque de 83%, et les non malades avec un pourcentage égale à 72,23%.

Conclusion générale

Ce mémoire avait pour ambition, au départ, de construire des modèles de classification qui aideront à la détection du cancer du sein à l'aide des facteurs du risque des femmes algériennes qui passeront le dépistage du cancer du sein au niveau du Centre Pierre et Marie Curie (CPMC) et d'élaborer des modèles qui permettront de calculer le risque de développer un cancer du sein d'une femme algérienne, mais d'après plusieurs mois de recherche pour passer notre stage pratique et pour collecter les informations nécessaires pour notre études , nous n'avons pas pu passer notre stage ni de collecter les données nécessaires pour élaborer nos modèles, ce qui nous a poussé à faire une enquête en ligne, et de travailler sur des données collectées à partir des associations françaises (Pour des raisons liées à l'absence des groupes d'associations algériennes sur les réseaux sociaux).

Après la réalisation de notre enquête en ligne, nous avons pu collecter 247 réponses (181 femmes malades et 66 femmes non malades), ce qui nous a permet de construire nos modèles, un arbre de décision avec une précision de 84% et un deuxième arbre de décision avec 72,5% de précision, où nous avons équilibré entre les proportions des femmes malades et non malades. Et un classifieur bayésien naïf gaussien basé sur le calcul des probabilités conditionnelles, avec un taux d'erreur de 20%. Ce qui nous donne un ensemble de performances acceptables et ces performances peuvent être améliorer en utilisant une base de données beaucoup plus volumineuse et en réalisant une enquête sur le terrain pour bien cibler la population.

Pour les résultats obtenus, nous avons trouvé en premier lieu que l'âge est le facteur de risque le plus discriminant de notre échantillon ce qui confirmera une partie de notre première hypothèse, et pour les 247 femmes nous avons trouvés que l'allaitement au sein est un facteur protecteur et l'allaitement avec une période plus d'un an diminuera le risque de développer un cancer du sein, et le risque augmentera si l'IMC est élevé.

Pour les modèles obtenus, les performances de la classification par arbre de décision (84% de précision) est meilleur de celle par classifieur bayésien naïf (80% de précision), ce qui confirmera notre troisième hypothèse.

Pour conclure, nous allons citer les points qui peuvent améliorer les résultats obtenus de notre étude et que nous aurons aimé travailler sur ou bien toucher :

- Travailler sur un grand nombre d'individus pour améliorer les performances et d'obtenir des modèles précis, car le Data mining nécessite des bases de données volumineuses.

- Faire une étude sur le terrain en collectant les caractéristiques (qui concerne les facteurs de risque du cancer du sein) sur des femmes algériennes au moment où elles passent le dépistage dans le CPMC, pour construire des modèles qui aident à la détection de cette maladie chez la femme algérienne et de la sensibilisée pour passer le dépistage, pour la détection précoce de cette maladie, afin de réduire la mortalité due au cancer du sein et son incidence en Algérie.

- Travailler avec un cancérologue qui nous aide à choisir les facteurs de risque du cancer du sein les plus importants pour la femme algérienne en nous orientant sur le choix des questions qui servent à la construction de notre base de données pour améliorer la précision de nos modèles.

Bibliographie

❖ Les ouvrages

Alex Campbell, Data Visualization Guide: Clear Introduction to Data Mining, Analysis, and Visualization, Kindle Edition (2021).

Fathalrahman Adam & Fathalrahman Adam, Knowledge Discovery in Big Data from Astronomy and Earth Observation : AstroGeoInformatics, Elsevier , 2020.

Herbert A. Edelstein, Introduction to Data Mining and Knowledge Discovery, Third Edition, 1999.

JAMES N. PARKER, M.D. ET PHILIP M. PARKER, PH. D, Breast Cancer: A Bibliography and Dictionary for Physicians, Patients, and Genome Researchers, 2007.

Jean-Claude Fournier, Théorie des graphes et applications : avec exercices et problèmes, 2ème édition revue et augmentée.

Jean-François Morère, Matti S.Aapro, Frédérique Penault-Llorca, Rémy Salmon, Le cancer du sein, Springer Paris , 2007.

Jean-Marc Classe, Cancer du sein : Dépistage et prise en charge, Elsevier Masson, (2016).

Max Bramer, Principles of Data Mining, Fourth Edition, 2020.

Sholom M. Weiss & Nitin Indurkhya, Predictive Data Mining : A Practical Guide, First Edition.

Stéphane Tufféry, Data Mining et statistique décisionnelle : La science des données, 5ème édition.

❖ Les articles et rapport

Alain-Jacques Valleron, L'épidémiologie humaine Conditions de son développement en France, et rôle des mathématiques, Académie des sciences : Rapport Science et Technologie, 2006.

André Nkondjock, Parviz Ghadirian, Facteurs de risque du cancer du sein, M/S : médecine sciences, Volume 21, numéro 2, 2005.

Djamel Abdelkader ZIGHED & Ricco RAKOTOMALALA, Extraction des Connaissances à partir des Données (ECD), Data Mining.

Épidémiologie des cancers du sein de la femme jeune en Afrique du Nord.

Institut national du cancer (France), guide d'information : Comprendre le cancer du sein, 2007.

Institut national du cancer (France), guide patients (J'ai un cancer : comprendre et être aidé), 2020.

LOUNICI MOSBAH.N, Data-Mining et apprentissage, Chapitre II : Modèle des arbres de décision.

Nahla Ben Amor, Salem Benferhat, Zied Elouedi, Réseaux bayésiens naïfs et arbres de décision dans les systèmes de détection d'intrusions, 2006.

Plan national : CANCER (2015-2019), Algérie, 2014.

Ricco RAKOTOMALALA, Arbres de décision, Revue MODULAD, 2005.

Usama Fayyad, Gregory Piatetsky-Shapiro & Padhraic Smyth, Knowledge Discovery and Data Mining: Towards a Unifying Framework, 1996.

❖ Les thèses et mémoires

BOUDHEB Tarik, Thèse de doctorat: Privacy Preserving Classification of Biomedical Data, Université Djilali Liabes Sidi BEL ABBES, 2018/2019.

FOUCAUT Aude-Marie, Thèse de doctorat : L'Activité Physique Adaptée en sénologie : des preuves scientifiques à la mise en œuvre de programmes auprès des patientes atteintes de cancer du sein, Université de Lyon, 2013.

Massinissa Saoudi, Thèse de doctorat : Conception of a wireless sensor network for decision making based on Data mining methods, Université de Bretagne occidentale - Brest, 2017.

❖ Les sources électroniques

Europa donna France- Association contre le cancer du sein, les différents types de cancer du sein. Récupéré sur http://www.europadonna.fr/le-cancer-du-sein/le-cancer-du-sein/differents-types-de-cancer/

Github, Introduction à l'apprentissage automatique.Récupéré sur https://projeduc.github.io/intro_apprentissage_automatique/bayes.html

Institut national du cancer (France), comprendre prévenir et dépister, types et stades des cancers. Récupéré sur https://www.e-cancer.fr/Comprendre-prevenir-depister/Qu-est-ce-qu-un-cancer/Types-et-stades-des-cancers#toc-types-de-cancers

Le CNAM, cours : Arbres de décision. Récupéré sur https://cedric.cnam.fr/vertigo/Cours/ml2/coursArbresDecision.html

Le journal des femmes Santé, Examens, Sein : anatomie, examens et maladies. Récupéré sur https://sante.journaldesfemmes.fr/fiches-anatomie-et-examens/2571039-sein-anatomie-examens-et-maladies/

National cancer institute (NIH), About cancer. Récupéré sur https://www.cancer.gov/about-cancer/understanding/what-is-cancer

Ooreka santé, Types de cancer du sein. Récupéré sur https://cancer-du-sein.ooreka.fr/comprendre/types-cancer-du-sein

Organisation des nations unies (ONU) : ONU info, l'actualité mondiale. Récupéré sur https://news.un.org/fr/story/2021/02/1088502

Organisation modiale de la santé, Thème de santé: Cancer. Récupéré sur https://www.who.int/topics/cancer/fr/

Organisation mondiale de la santé, centre des médias, principaux repères, détail, cancer. Récupéré sur https://www.who.int/fr/news-room/fact-sheets/detail/cancer

Société canadienne du cancer, information sur le cancer : diagnostic et traitement. Récupéré sur https://www.cancer.ca/fr-ca/cancer-information/diagnosis-and-treatment/tests-and-procedures/fine-needle-aspiration/?region=qc

Swakinome, difference between KDD and DM. Récupéré sur https://fr.sawakinome.com/articles/database/difference-between-kdd-and-data-mining.html

Annexes

Annexe 1 : Questionnaire version en ligne.

Enquête sur les facteurs de risque de cancer du sein

En étant la première cause de mortalité et le cancer le plus fréquent chez la femme dans le monde, ce questionnaire a été élaboré dans le cadre de la réalisation d'un mémoire de fin d'étude, afin de collecter des informations sur les facteurs de risque de cette maladie, et ces questions-là vous concerne en étant une femme, soit atteinte du cancer de sein ou bien non.
Merci d'avance pour votre temps ! et sachez que vos réponses ne seront traitées qu'à des fins statistiques et de manière strictement anonyme.

*Obligatoire

1. Quel est votre âge ? *

2. Quel est votre poids ? (En kilogrammes) *

3. Quel est votre taille ? (En centimètres) *

4. Avez-vous un cancer du sein ? *

Une seule réponse possible.

- ◯ Oui
- ◯ Non

5. Avez-vous un membre de votre famille atteint d'un cancer du sein ? *

Une seule réponse possible.

- ◯ Oui
- ◯ Non

6. Si oui précisez (Mère, sœur, grand-mère, cousine, tente…) :

Plusieurs réponses possibles.

- ☐ Grand-mère
- ☐ Mère
- ☐ Tante
- ☐ Sœur
- ☐ Cousine
- Autre : ☐ _____

7. Quelle est votre situation matrimoniale ? *

Une seule réponse possible.

- ◯ Célibataire
- ◯ Mariée
- ◯ Divorcée
- ◯ Veuve

8. Parité : Avez-vous des enfants ? *

Une seule réponse possible.

- ◯ Oui
- ◯ Non

9. Si oui, combien ?

_____

10. Age à la première naissance : A quel âge avez-vous votre premier enfant ?

_____

11. Allaitement : Avez-vous allaitez vos enfants au sein ?

Une seule réponse possible.

- ◯ Oui
- ◯ Non

12. Si oui, quelle est la durée de votre allaitement ?

Une seule réponse possible.

- ◯ Moins de 6 mois
- ◯ Entre 6 mois et un an
- ◯ Plus d'un an

13. L'âge aux premières règles : à quel âge vous avez eu vos premières règles ? *

_____

14. L'âge à la ménopause :

_____

15. Suivez-vous un traitement hormonal de la ménopause ?

Une seule réponse possible.

- ◯ Oui
- ◯ Non

16. Prenez-vous des pilules contraceptives ? *

Une seule réponse possible.

- ◯ Oui
- ◯ Non

17. Pratiquez-vous une activité physique ? *

Une seule réponse possible.

- ◯ Oui
- ◯ Non

18. Si oui, combien de fois par semaine ?

_____

Annexe 2 : Tableau de données.

Id	X1	X2	X3	X4	Y	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19
1	52	98	1,56	40,2695595	Oui	Non	0	Mariée	Oui	1	<30	Non	Pas d'allaitement	12	Oui	>=50	Non	Oui	Non	0
2	63	86	1,7	29,75778547	Oui	Non	0	Divorcée	Oui	2	<30	Oui	Moins d'un an	13	Oui	>=50	Non	Oui	Oui	3
3	55	78	1,68	27,63605442	Non	Non	0	Veuve	Oui	1	>=30	Non	Pas d'allaitement	13	Oui	>=50	Non	Oui	Oui	1
4	33	85	1,65	31,22130395	Oui	Oui	2	Mariée	Oui	2	<30	Oui	Moins d'un an	12	Non	Non ménopausée	Non	Oui	Non	0
5	44	80	1,68	28,3446712	Oui	Oui	1	Mariée	Oui	3	<30	Non	Pas d'allaitement	13	Non	Non ménopausée	Non	Oui	Non	0
6	39	90	1,68	31,8877551	Oui	Non	0	Mariée	Oui	1	<30	Non	Pas d'allaitement	14	Non	Non ménopausée	Non	Oui	Oui	7
7	53	78	1,56	32,05128205	Oui	Non	0	Veuve	Oui	2	<30	Oui	Moins d'un an	12	Oui	>=50	Oui	Oui	Non	0
8	45	57	1,6	22,265625	Oui	Oui	1	Mariée	Oui	1	>=30	Non	Pas d'allaitement	12	Non	Non ménopausée	Non	Oui	Non	0
9	54	59	1,67	21,1552942	Oui	Oui	1	Divorcée	Oui	1	>=30	Non	Pas d'allaitement	14	Oui	>=50	Non	Non	Oui	7
10	65	52	1,6	20,3125	Non	Non	0	Divorcée	Oui	1	<30	Non	Pas d'allaitement	13	Oui	>=50	Non	Oui	Non	0
11	46	61	1,66	22,13673973	Oui	Non	0	Divorcée	Oui	3	<30	Oui	Moins d'un an	10	Oui	<50	Non	Oui	Oui	3
12	43	52	1,64	19,33372992	Oui	Oui	1	Mariée	Oui	2	<30	Non	Pas d'allaitement	10,5	Non	Non ménopausée	Non	Oui	Oui	2
13	61	82	1,6	32,03125	Oui	Oui	1	Mariée	Oui	2	<30	Non	Pas d'allaitement	12	Oui	>=50	Non	Oui	Oui	7
14	29	97	1,87	27,73885441	Oui	Non	0	Mariée	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	12	Non	Non ménopausée	Non	Oui	Non	0
15	39	67	1,63	25,21735858	Oui	Non	0	Veuve	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	14	Oui	>=50	Non	Non	Non	0
16	41	54	1,67	19,36247266	Non	Non	0	Mariée	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	12	Non	Non ménopausée	Non	Oui	Non	0
17	39	82	1,55	34,13111342	Oui	Non	0	Veuve	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	13	Non	Non ménopausée	Non	Non	Non	0
18	40	58	1,67	20,79672989	Oui	Non	0	Mariée	Oui	2	<30	Oui	Moins d'un an	11	Non	Non ménopausée	Non	Non	Non	0
19	61	60	1,7	20,76124567	Oui	Oui	1	Divorcée	Oui	1	<30	Non	Pas d'allaitement	14	Oui	>=50	Non	Non	Oui	6
20	39	65	1,68	23,03004535	Oui	Non	0	Célibataire	Oui	2	<30	Oui	Moins d'un an	14	Non	Non ménopausée	Oui	Non	Oui	7
21	42	64	1,7	22,14532872	Oui	Non	0	Mariée	Oui	2	<30	Oui	Moins d'un an	14	Non	Non ménopausée	Non	Oui	Oui	1
22	42	60	1,65	22,03856749	Non	Oui	1	Célibataire	Oui	3	<30	Non	Pas d'allaitement	10	Non	Non ménopausée	Non	Non	Non	0

23	42	50	1,57	20,28479857	Oui	Oui	1	Divorcée	Oui	3	<30	Oui	Moins d'un an	14	Non	Non ménopausée	Non	Non	Non	0
24	43	46	1,62	17,52781588	Oui	Non	0	Célibataire	Oui	1	>=30	Non	Pas d'allaitement	15	Oui	<50	Non	Non	Oui	1
25	52	90	1,66	32,66076354	Oui	Non	0	Divorcée	Oui	2	>=30	Oui	Plus d'un an	11	Oui	>=50	Non	Non	Oui	1
26	48	100	1,65	36,73094582	Oui	Oui	2	Mariée	Oui	4	<30	Oui	Moins d'un an	15	Non	Non ménopausée	Non	Non	Oui	2
27	59	67	1,67	24,02380867	Non	Oui	1	Célibataire	Oui	3	<30	Oui	Moins d'un an	12	Oui	>=50	Non	Non	Oui	4
28	52	63	1,72	21,29529475	Oui	Non	0	Mariée	Oui	2	>=30	Oui	Moins d'un an	14	Oui	<50	Oui	Non	Non	0
29	40	53	1,65	19,46740129	Oui	Non	0	Mariée	Oui	3	<30	Non	Pas d'allaitement	14	Non	Non ménopausée	Non	Oui	Non	0
30	56	64	1,6	25	Oui	Oui	2	Divorcée	Oui	1	<30	Non	Pas d'allaitement	14	Oui	>=50	Non	Oui	Non	0
31	48	60	1,65	22,03856749	Oui	Non	0	Mariée	Oui	3	<30	Non	Pas d'allaitement	12	Oui	<50	Oui	Oui	Oui	2
32	43	59	1,6	23,046875	Oui	Non	0	Divorcée	Oui	2	<30	Non	Pas d'allaitement	14	Non	Non ménopausée	Oui	Non	Non	0
33	42	111	1,7	38,4083045	Oui	Oui	2	Mariée	Oui	4	<30	Non	Moins d'un an	11	Non	Non ménopausée	Non	Oui	Oui	2

34	38	112	1,65	41,13865932	Oui	Non	0	Mariée	Oui	1	>=30	Oui	Moins d'un an	16	Non	Non ménopausée	Oui	Non	Non	0
35	43	68	1,58	27,23922448	Non	Oui	1	Mariée	Oui	7	<30	Oui	Plus d'un an	12	Non	Non ménopausée	Non	Non	Oui	4
36	73	88	1,65	32,32323232	Non	Non	0	Divorcée	Oui	0	Pas encore	Pas d'enfants	Pas d'allaitement	11	Oui	>=50	Non	Oui	Non	0
37	57	154	1,65	56,56565657	Oui	Oui	3	Mariée	Oui	3	<30	Oui	Moins d'un an	12	Oui	<50	Non	Non	Non	0
38	39	55	1,6	21,484375	Oui	Oui	1	Mariée	Oui	3	<30	Oui	Plus d'un an	11	Non	Non ménopausée	Non	Oui	Non	0
39	52	63	1,66	22,86253448	Oui	Non	0	Célibataire	Oui	2	>=30	Oui	Moins d'un an	16	Oui	>=50	Non	Non	Non	0
40	41	80	1,6	31,25	Oui	Oui	1	Mariée	Oui	2	>=30	Oui	Moins d'un an	12	Non	Non ménopausée	Non	Oui	Oui	2
41	46	48	1,52	20,77562327	Oui	Non	0	Mariée	Oui	3	<30	Oui	Plus d'un an	11	Non	Non ménopausée	Non	Non	Oui	4
42	53	59	1,67	21,1552942	Oui	Non	0	Divorcée	Oui	1	<30	Non	Pas d'allaitement	11	Non	Non ménopausée	Non	Oui	Oui	1
43	43	65	1,65	23,87511478	Oui	Oui	1	Divorcée	Oui	1	<30	Non	Pas d'allaitement	14	Oui	<50	Non	Non	Non	0
44	33	60	1,63	22,58270917	Non	Non	0	Mariée	Oui	2	<30	Oui	Plus d'un an	11	Non	Non ménopausée	Non	Non	Non	0
45	39	51	1,78	16,09645247	Oui	Oui	1	Mariée	Oui	1	>=30	Non	Pas d'allaitement	13	Non	Non ménopausée	Non	Oui	Non	0
46	52	67	1,72	22,64737696	Oui	Oui	2	Célibataire	Oui	1	>=30	Non	Pas d'allaitement	13	Non	Non ménopausée	Non	Non	Non	0

47	32	67	1,59	26,50211621	Oui	Oui	1	Célibataire	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	11	Non	Non ménopausée	Non	Oui	Oui	3
48	39	64	1,71	21,88707637	Oui	Non	0	Mariée	Oui	3	<30	Non	Pas d'allaitement	12	Non	Non ménopausée	Non	Non	Oui	1
49	47	57	1,71	19,49317739	Oui	Oui	1	Mariée	Oui	2	<30	Oui	Moins d'un an	14	Oui	<50	Non	Non	Oui	3
50	43	57	1,73	19,04507334	Oui	Non	0	Mariée	Oui	2	>=30	Oui	Moins d'un an	14	Non	Non ménopausée	Non	Oui	Oui	3
51	47	60	1,67	21,51385851	Oui	Oui	2	Mariée	Oui	2	>=30	Non	Pas d'allaitement	14	Non	Non ménopausée	Non	Non	Non	0
52	45	47	1,73	15,7038324	Oui	Oui	1	Divorcée	Oui	4	<30	Oui	Moins d'un an	15	Oui	<50	Non	Non	Oui	3
53	41	79	1,69	27,66009594	Oui	Oui	3	Mariée	Oui	2	<30	Non	Pas d'allaitement	10	Non	Non ménopausée	Non	Oui	Non	0
54	50	72	1,65	26,44628099	Oui	Non	0	Mariée	Oui	1	>=30	Non	Pas d'allaitement	11	Oui	<50	Non	Non	Oui	5
55	48	73	1,68	25,86451247	Oui	Oui	2	Mariée	Oui	2	<30	Oui	Moins d'un an	12,5	Oui	<50	Oui	Non	Non	0
56	61	82	1,63	30,86303587	Oui	Non	0	Célibataire	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	11	Oui	>=50	Non	Oui	Non	0
57	58	61	1,7	21,10726644	Non	Non	0	Mariée	Oui	1	<30	Non	Pas d'allaitement	10	Oui	<50	Non	Non	Non	0
58	58	91	1,7	31,48788927	Oui	Oui	1	Mariée	Oui	3	>=30	Oui	Moins d'un an	14	Oui	>=50	Non	Oui	Non	0
59	35	76	1,75	24,81632653	Oui	Non	0	Célibataire	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	12	Non	Non ménopausée	Non	Oui	Oui	4
60	29	50	1,6	19,53125	Non	Oui	1	Mariée	Oui	2	<30	Non	Pas d'allaitement	14	Non	Non ménopausée	Non	Non	Non	0
61	47	115	1,65	42,2405877	Oui	Oui	1	Mariée	Oui	1	<30	Oui	Moins d'un an	12	Non	Non ménopausée	Non	Non	Oui	2
62	50	55	1,67	19,72103697	Oui	Non	0	Mariée	Oui	2	<30	Non	Pas d'allaitement	15	Non	Non ménopausée	Non	Non	Non	0
63	72	73	1,64	27,14158239	Oui	Non	0	Mariée	Oui	1	<30	Oui	Moins d'un an	13	Oui	<50	Non	Non	Oui	2
64	41	69	1,69	24,15881797	Oui	Oui	2	Mariée	Oui	3	<30	Oui	Moins d'un an	12	Non	Non ménopausée	Non	Non	Non	0
65	56	69	1,65	25,34435262	Oui	Non	0	Divorcée	Oui	1	>=30	Oui	Moins d'un an	10	Oui	>=50	Non	Non	Oui	2
66	52	83	1,685	29,23332952	Oui	Oui	1	Mariée	Oui	3	<30	Oui	Plus d'un an	15	Oui	<50	Non	Non	Oui	2
67	42	86	1,63	32,36854981	Oui	Oui	1	Mariée	Oui	2	<30	Non	Pas d'allaitement	12	Non	Non ménopausée	Non	Non	Oui	2

68	37	82	1,66	29,75758456	Oui	Non	0	Mariée	Oui	3	<30	Non	Pas d'allaitement	14	Non	Non ménopausée	Non	Non	Oui	2
69	49	63	1,61	24,30461788	Oui	Non	0	Mariée	Oui	3	<30	Oui	Plus d'un an	12	Oui	<50	Non	Non	Non	0
70	33	75	1,65	27,54820937	Oui	Non	0	Mariée	Oui	2	<30	Oui	Moins d'un an	15	Non	Non ménopausée	Non	Non	Oui	3

71	54	60	1,57	24,34175829	Oui	Oui	2	Célibataire	Oui	1	<30	Non	Pas d'allaitement	10	Oui	<50	Non	Non	Non	0
72	35	58	1,62	22,10028959	Oui	Oui	1	Mariée	Oui	2	<30	Non	Pas d'allaitement	12	Non	Non ménopausée	Non	Oui	Non	0
73	32	52	1,62	19,81405274	Oui	Non	0	Mariée	Oui	2	<30	Oui	Plus d'un an	12	Non	Non ménopausée	Non	Oui	Oui	2
74	48	92	1,65	33,79247016	Oui	Non	0	Mariée	Oui	3	<30	Oui	Moins d'un an	14	Non	Non ménopausée	Non	Non	Non	0
75	52	80	1,72	27,04164413	Oui	Oui	1	Mariée	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	13	Oui	<50	Non	Non	Non	0
76	55	80	1,66	29,03178981	Oui	Oui	2	Mariée	Oui	2	<30	Non	Pas d'allaitement	14	Oui	>=50	Non	Oui	Oui	3
77	59	59	1,57	23,93606231	Oui	Non	0	Divorcée	Oui	2	<30	Oui	Moins d'un an	12	Oui	>=50	Non	Oui	Oui	2
78	42	90	1,7	31,14186851	Oui	Non	0	Célibataire	Oui	5	>=30	Oui	Moins d'un an	9	Non	Non ménopausée	Non	Non	Oui	1
79	52	90	1,69	31,5115017	Oui	Oui	1	Mariée	Oui	1	<30	Non	Pas d'allaitement	14	Non	Non ménopausée	Oui	Non	Oui	3
80	52	59	1,7	20,41522491	Oui	Non	0	Veuve	Oui	3	<30	Non	Pas d'allaitement	13	Oui	>=50	Non	Non	Non	0
81	51	78	1,72	26,36560303	Oui	Oui	1	Mariée	Oui	2	<30	Oui	Moins d'un an	15	Non	Non ménopausée	Non	Non	Oui	1
82	43	86	1,68	30,47052154	Oui	Non	0	Mariée	Oui	7	>=30	Non	Pas d'allaitement	12	Non	Non ménopausée	Non	Non	Non	0
83	47	95	1,7	32,87197232	Oui	Non	0	Mariée	Oui	3	<30	Oui	Moins d'un an	13	Non	Non ménopausée	Non	Oui	Oui	2
84	39	55	1,7	19,03114187	Oui	Non	0	Mariée	Oui	2	>=30	Oui	Moins d'un an	11	Non	Non ménopausée	Non	Non	Non	0
85	30	80	1,75	26,12244898	Oui	Oui	1	Mariée	Oui	1	<30	Oui	Moins d'un an	10	Non	Non ménopausée	Oui	Non	Non	0
86	57	80	1,59	31,64431787	Oui	Non	0	Mariée	Oui	2	<30	Oui	Moins d'un an	13	Non	Non ménopausée	Non	Non	Non	0
87	57	64	1,77	20,42835711	Oui	Oui	3	Célibataire	Oui	3	<30	Oui	Moins d'un an	12	Oui	>=50	Non	Non	Oui	4
88	36	60	1,69	21,0076678	Oui	Oui	1	Célibataire	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	12	Non	Non ménopausée	Non	Oui	Oui	2
89	54	80	1,57	32,45567772	Oui	Oui	2	Divorcée	Oui	3	<30	Oui	Moins d'un an	11	Non	Non ménopausée	Non	Oui	Non	0
90	49	102	1,71	34,88252796	Oui	Oui	1	Mariée	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	13,5	Oui	<50	Non	Non	Non	0
91	52	50	1,58	20,02884153	Oui	Oui	2	Divorcée	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	14,5	Oui	<50	Non	Non	Oui	3
92	47	76	1,68	26,92743764	Oui	Oui	2	Mariée	Oui	6	<30	Non	Pas d'allaitement	11	Oui	<50	Non	Non	Oui	3
93	52	72	1,63	27,09925101	Oui	Non	0	Mariée	Oui	5	<30	Oui	Plus d'un an	13	Non	Non ménopausée	Non	Non	Oui	7
94	55	94	1,72	31,77393186	Oui	Non	0	Divorcée	Oui	4	>=30	Oui	Moins d'un an	14	Oui	>=50	Non	Non	Non	0
95	46	62	1,65	22,77318641	Oui	Non	0	Mariée	Oui	3	<30	Oui	Moins d'un an	11	Non	Non ménopausée	Non	Non	Oui	2

96	55	85	1,59	33,62208773	Oui	Oui	1	Veuve	Oui	2	>=30	Oui	Moins d'un an	10	Oui	>=50	Non	Non	Oui	7
97	41	59	1,68	20,90419501	Oui	Oui	1	Mariée	Oui	1	>=30	Oui	Moins d'un an	14	Non	Non ménopausée	Non	Non	Oui	5
98	40	74	1,72	25,01352082	Oui	Oui	3	Célibataire	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	14	Non	Non ménopausée	Non	Non	Oui	2
99	49	68	1,66	24,67702134	Oui	Oui	3	Divorcée	Oui	2	<30	Oui	Moins d'un an	12	Oui	<50	Non	Non	Oui	2
100	45	71	1,64	26,39797739	Oui	Oui	1	Mariée	Oui	2	<30	Non	Pas d'allaitement	10	Non	Non ménopausée	Non	Oui	Non	0
101	46	60	1,65	22,03856749	Oui	Oui	2	Mariée	Oui	2	>=30	Oui	Moins d'un an	14	Non	Non ménopausée	Non	Non	Oui	1

102	60	75	1,63	28,22838647	Oui	Oui	1	Mariée	Oui	1	>=30	Oui	Moins d'un an	14	Oui	>=50	Oui	Oui	Oui	7
103	46	77	1,63	28,98114344	Oui	Oui	2	Mariée	Oui	3	<30	Oui	Moins d'un an	12	Non	Non ménopausée	Non	Oui	Non	0
104	36	62	1,68	21,96712018	Oui	Non	0	Mariée	Oui	3	<30	Non	Pas d'allaitement	12	Non	Non ménopausée	Non	Oui	Oui	4
105	52	70	1,73	23,38868656	Oui	Oui	1	Divorcée	Oui	1	>=30	Oui	Plus d'un an	11,5	Non	Non ménopausée	Non	Non	Oui	2
106	41	76	1,69	26,60971255	Oui	Non	0	Mariée	Oui	3	<30	Oui	Moins d'un an	14	Non	Non ménopausée	Non	Non	Oui	1
107	41	61	1,65	22,40587695	Oui	Oui	1	Mariée	Oui	3	<30	Oui	Moins d'un an	15	Non	Non ménopausée	Non	Non	Non	0
108	32	62	1,71	21,20310523	Oui	Oui	1	Mariée	Oui	2	<30	Oui	Moins d'un an	14	Non	Non ménopausée	Non	Oui	Oui	2
109	68	75	1,72	25,35154137	Oui	Non	0	Veuve	Oui	2	<30	Oui	Moins d'un an	14	Oui	<50	Non	Non	Non	0
110	55	65	1,7	22,49134948	Oui	Oui	2	Mariée	Oui	1	>=30	Oui	Moins d'un an	15	Oui	>=50	Non	Oui	Oui	5
111	38	77	1,6	30,078125	Oui	Oui	1	Célibataire	Oui	3	<30	Oui	Plus d'un an	14	Non	Non ménopausée	Non	Non	Non	0
112	40	65	1,6	25,390625	Oui	Oui	1	Mariée	Oui	3	<30	Non	Pas d'allaitement	12	Non	Non ménopausée	Non	Non	Oui	1
113	37	75	1,68	26,57312925	Oui	Non	0	Mariée	Oui	1	>=30	Oui	Moins d'un an	12	Non	Non ménopausée	Non	Oui	Non	0
114	33	60	1,7	20,76124567	Oui	Oui	1	Mariée	Oui	1	>=30	Non	Pas d'allaitement	14	Non	Non ménopausée	Non	Non	Non	0
115	36	60	1,6	23,4375	Oui	Oui	1	Célibataire	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	13	Non	Non ménopausée	Non	Oui	Non	0
116	47	90	1,74	29,72651605	Non	Non	0	Mariée	Oui	4	<30	Oui	Moins d'un an	12	Non	Non ménopausée	Non	Oui	Non	0
117	54	55	1,65	20,2020202	Oui	Oui	1	Divorcée	Oui	3	<30	Non	Pas d'allaitement	14	Oui	<50	Non	Non	Non	0
118	41	55	1,6	21,484375	Oui	Oui	1	Mariée	Oui	2	>=30	Oui	Moins d'un an	12	Non	Non ménopausée	Non	Non	Non	0
119	57	70	1,5	31,11111111	Oui	Non	0	Mariée	Oui	2	<30	Non	Pas d'allaitement	12	Oui	>=50	Non	Oui	Non	0

120	54	150	1,65	55,09641873	Non	Non	0	Célibataire	Oui	7	<30	Oui	Moins d'un an	12	Oui	<50	Non	Non	Non	0
121	41	67	1,6	26,171875	Oui	Non	0	Mariée	Oui	3	<30	Oui	Moins d'un an	11	Non	Non ménopausée	Non	Non	Non	0
122	69	68	1,68	24,09297052	Oui	Oui	1	Mariée	Oui	3	<30	Oui	Moins d'un an	13	Oui	>=50	Non	Non	Oui	2
123	38	80	1,7	27,6816609	Non	Oui	1	Mariée	Oui	1	>=30	Non	Pas d'allaitement	11	Non	Non ménopausée	Non	Oui	Non	0
124	51	66	1,74	21,79944511	Oui	Oui	1	Mariée	Oui	2	<30	Oui	Moins d'un an	13	Non	Non ménopausée	Non	Oui	Non	0
125	66	110	1,72	37,18226068	Oui	Oui	1	Mariée	Oui	3	<30	Non	Pas d'allaitement	13	Oui	>=50	Non	Non	Non	0
126	40	140	1,65	51,42332415	Oui	Oui	1	Célibataire	Oui	2	<30	Oui	Plus d'un an	11	Oui	<50	Non	Non	Oui	7
127	31	69	1,61	26,61934339	Non	Non	0	Mariée	Oui	1	<30	Non	Pas d'allaitement	13	Non	Non ménopausée	Non	Oui	Oui	1
128	42	70	1,7	24,22145329	Oui	Non	0	Mariée	Oui	3	<30	Non	Pas d'allaitement	11	Non	Non ménopausée	Non	Non	Oui	1
129	45	76	1,7	26,29757785	Oui	Non	0	Divorcée	Oui	1	<30	Oui	Moins d'un an	11,5	Non	Non ménopausée	Non	Non	Oui	3
130	46	78	1,6	30,46875	Oui	Oui	1	Mariée	Oui	5	<30	Non	Pas d'allaitement	14	Non	Non ménopausée	Non	Non	Non	0
131	45	81	1,75	26,44897959	Oui	Non	0	Mariée	Oui	2	<30	Non	Pas d'allaitement	11	Non	Non ménopausée	Non	Non	Non	0
132	52	65	1,6	25,390625	Oui	Non	0	Mariée	Oui	4	<30	Oui	Moins d'un an	14	Oui	>=50	Non	Non	Non	0
133	35	70	1,65	25,71166208	Non	Non	0	Célibataire	Oui	1	>=30	Oui	Moins d'un an	15	Non	Non ménopausée	Non	Oui	Non	0
134	43	66	1,73	22,05219018	Oui	Oui	1	Célibataire	Oui	2	<30	Non	Pas d'allaitement	15	Non	Non ménopausée	Non	Non	Non	0
135	57	57	1,69	19,95728441	Oui	Non	0	Mariée	Oui	3	<30	Non	Pas d'allaitement	13	Oui	<50	Oui	Non	Oui	2

136	50	54	1,63	20,32443826	Oui	Oui	1	Mariée	Oui	1	<30	Oui	Moins d'un an	12	Non	Non ménopausée	Non	Non	Oui	5
137	56	90	1,65	33,05785124	Oui	Oui	1	Célibataire	Non	0	Pas encore	Non	Pas d'allaitement	14	Oui	>=50	Non	Non	Non	0
138	22	50	1,64	18,59012493	Non	Oui	1	Célibataire	Non	0	Pas encore	Non	Pas d'allaitement	12	Non	Non ménopausée	Non	Non	Oui	3
139	43	70	1,7	24,22145329	Oui	Non	0	Mariée	Oui	4	<30	Oui	Moins d'un an	13	Non	Non ménopausée	Non	Non	Non	0
140	60	60	1,6	23,4375	Oui	Non	0	Mariée	Oui	5	>=30	Oui	Plus d'un an	12	Oui	<50	Non	Non	Non	0
141	54	90	1,68	31,8877551	Oui	Non	0	Mariée	Oui	5	<30	Oui	Moins d'un an	14	Oui	>=50	Non	Oui	Non	0
142	25	58	1,59	22,94213045	Non	Oui	1	Célibataire	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	13	Non	Non ménopausée	Non	Non	Oui	2
143	43	65	1,72	21,97133586	Non	Non	0	Mariée	Oui	2	<30	Oui	Moins d'un an	11	Non	Non ménopausée	Non	Oui	Non	0

144	48	87	1,74	28,73563218	Non	Non	0	Mariée	Oui	3	<30	Oui	Plus d'un an	12	Non	Non ménopausée	Non	Non	Non	0
145	40	80	1,69	28,01022373	Non	Non	0	Mariée	Oui	2	<30	Oui	Moins d'un an	14	Non	Non ménopausée	Non	Non	Non	0
146	18	50	1,58	20,02884153	Non	Oui	1	Célibataire	Non	0	Pas encore	Non	Pas d'allaitement	11	Non	Non ménopausée	Non	Non	Oui	3
147	24	80	1,67	28,68514468	Non	Non	0	Mariée	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	11	Non	Non ménopausée	Non	Non	Non	0
148	29	77	1,68	27,28174603	Non	Oui	1	Célibataire	Non	0	Pas encore	Non	Pas d'allaitement	13	Non	Non ménopausée	Non	Non	Non	0
149	60	78	1,73	26,06167931	Non	Oui	1	Veuve	Oui	3	>=30	Oui	Moins d'un an	15	Oui	<50	Non	Non	Non	0
150	19	78	1,61	30,09143166	Non	Oui	1	Célibataire	Non	0	Pas encore	Non	Pas d'allaitement	10	Non	Non ménopausée	Non	Non	Non	0
151	47	60	1,68	21,2585034	Oui	Oui	1	Mariée	Oui	5	<30	Oui	Moins d'un an	14	Non	Non ménopausée	Non	Non	Oui	3
152	38	81	1,7	28,02768166	Oui	Non	0	Mariée	Oui	4	<30	Oui	Moins d'un an	11	Non	Non ménopausée	Non	Non	Oui	5
153	38	61	1,68	21,61281179	Oui	Non	0	Mariée	Oui	2	>=30	Oui	Moins d'un an	11	Non	Non ménopausée	Oui	Non	Oui	1
154	47	63	1,7	21,79930796	Oui	Non	0	Célibataire	Oui	2	<30	Non	Pas d'allaitement	16	Non	Non ménopausée	Non	Non	Non	0
155	55	80	1,66	29,03178981	Oui	Non	0	Divorcée	Oui	3	<30	Oui	Moins d'un an	15	Oui	>=50	Non	Non	Oui	7
156	42	52	1,69	18,20664543	Oui	Oui	1	Mariée	Oui	2	>=30	Oui	Moins d'un an	14	Non	Non ménopausée	Non	Non	Oui	2
157	50	70	1,6	27,34375	Oui	Non	0	Mariée	Oui	4	<30	Non	Pas d'allaitement	14	Non	Non ménopausée	Oui	Non	Oui	5
158	44	84	1,7	29,06574394	Oui	Non	0	Mariée	Oui	2	>=30	Oui	Moins d'un an	14	Non	Non ménopausée	Non	Non	Oui	1
159	66	80	1,65	29,38475666	Oui	Oui	1	Mariée	Oui	2	<30	Oui	Moins d'un an	11	Oui	>=50	Non	Non	Non	0
160	35	63	1,56	25,88757396	Oui	Non	0	Mariée	Oui	2	<30	Non	Pas d'allaitement	14	Non	Non ménopausée	Non	Non	Oui	2
161	46	72	1,8	22,22222222	Oui	Oui	2	Mariée	Oui	2	<30	Oui	Moins d'un an	14	Non	Non ménopausée	Non	Non	Oui	7
162	53	86	1,67	30,83653053	Oui	Non	0	Célibataire	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	13	Oui	>=50	Non	Non	Oui	5
163	50	152	1,57	61,66578766	Oui	Oui	2	Mariée	Oui	3	<30	Non	Pas d'allaitement	13	Non	Non ménopausée	Non	Non	Non	0
164	44	75	1,65	27,54820937	Oui	Non	0	Mariée	Oui	5	<30	Non	Pas d'allaitement	13	Non	Non ménopausée	Non	Oui	Non	0
165	58	73	1,7	25,25951557	Oui	Non	0	Mariée	Oui	5	<30	Non	Pas d'allaitement	13	Oui	<50	Non	Non	Non	0
166	45	57	1,6	22,265625	Oui	Non	0	Célibataire	Oui	1	<30	Non	Pas d'allaitement	12	Non	Non ménopausée	Non	Non	Non	0
167	64	90	1,68	31,8877551	Oui	Non	0	Divorcée	Oui	2	<30	Oui	Moins d'un an	13	Oui	>=50	Non	Non	Oui	7
168	39	83	1,68	29,40759637	Oui	Non	0	Mariée	Oui	5	<30	Oui	Moins d'un an	12	Non	Non ménopausée	Non	Oui	Non	0

169	20	53	1,69	18,55677322	Non	Non	0	Mariée	Oui	1	<30	Pas d'enfants	Moins d'un an	8	Non	Non ménopausée	Non	Oui	Non	0

170	20	55	1,58	22,03172568	Non	Non	0	Célibataire	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	11	Non	Non ménopausée	Non	Non	Oui	2
171	45	70	1,64	26,0261749	Non	Non	0	Mariée	Oui	2	<30	Oui	Moins d'un an	12	Non	Non ménopausée	Non	Non	Non	0
172	52	58	1,7	20,06920415	Oui	Non	0	Veuve	Oui	3	<30	Non	Pas d'allaitement	13	Oui	>=50	Non	Non	Non	0
173	47	58	1,69	20,30741221	Oui	Oui	2	Divorcée	Oui	4	<30	Non	Pas d'allaitement	16	Non	Non ménopausée	Non	Non	Oui	3
174	35	70	1,58	28,04037814	Oui	Oui	2	Célibataire	Oui	1	<30	Non	Pas d'allaitement	13	Oui	<50	Oui	Non	Oui	3
175	38	95	1,67	34,06360931	Oui	Oui	1	Mariée	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	12	Oui	<50	Non	Non	Oui	2
176	62	43	1,55	17,89802289	Non	Non	0	Mariée	Oui	3	<30	Oui	Moins d'un an	13	Oui	>=50	Non	Non	Oui	5
177	40	51	1,6	19,921875	Non	Non	0	Mariée	Oui	3	>=30	Oui	Moins d'un an	15	Non	Non ménopausée	Non	Non	Non	0
178	35	53	1,69	18,55677322	Oui	Non	0	Mariée	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	13	Non	Non ménopausée	Non	Non	Non	0
179	32	49	1,58	19,6282647	Non	Oui	1	Célibataire	Oui	1	<30	Non	Pas d'allaitement	14	Non	Non ménopausée	Non	Oui	Non	0
180	37	59	1,6	23,046875	Non	Oui	1	Mariée	Oui	2	>=30	Non	Pas d'allaitement	13	Non	Non ménopausée	Non	Non	Non	0
181	28	56	1,73	18,71094925	Non	Oui	2	Célibataire	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	11	Non	Non ménopausée	Non	Non	Oui	1
182	39	80	1,65	29,38475666	Non	Non	0	Mariée	Oui	1	<30	Oui	Moins d'un an	12	Non	Non ménopausée	Non	Non	Oui	2
183	56	66	1,65	24,24242424	Oui	Oui	3	Mariée	Oui	2	<30	Non	Pas d'allaitement	11	Oui	>=50	Non	Non	Oui	7
184	64	52	1,73	17,37445287	Non	Non	0	Divorcée	Oui	2	>=30	Oui	Moins d'un an	12	Oui	>=50	Non	Non	Non	0
185	39	81	1,7	28,02768166	Non	Oui	2	Mariée	Oui	2	<30	Oui	Plus d'un an	14	Non	Non ménopausée	Non	Non	Oui	3
186	45	49	1,73	16,37208059	Oui	Non	0	Divorcée	Oui	4	<30	Oui	Moins d'un an	15	Oui	<50	Non	Non	Oui	3
187	48	68	1,7	23,52941176	Oui	Oui	1	Divorcée	Oui	1	>=30	Non	Pas d'allaitement	15	Oui	<50	Non	Non	Non	0
188	39	68	1,58	27,23922448	Oui	Non	0	Célibataire	Oui	1	<30	Oui	Moins d'un an	12	Non	Non ménopausée	Non	Non	Oui	7
189	60	72	1,68	25,51020408	Oui	Oui	1	Mariée	Oui	2	<30	Oui	Moins d'un an	15	Oui	<50	Non	Non	Non	0
190	46	59	1,57	23,93606231	Oui	Non	0	Mariée	Oui	4	<30	Oui	Moins d'un an	13	Non	Non ménopausée	Non	Non	Oui	3
191	49	69	1,67	24,74093729	Oui	Non	0	Mariée	Oui	2	<30	Non	Pas d'allaitement	11	Non	Non ménopausée	Non	Non	Non	0
192	48	80	1,75	26,12244898	Oui	Oui	1	Mariée	Oui	4	<30	Non	Pas d'allaitement	11	Non	Non ménopausée	Non	Non	Oui	5

193	26	46	1,56	18,90203813	Non	Non	0	Célibataire	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	11	Non	Non ménopausée	Non	Non	Oui	3
194	57	65	1,66	23,58832922	Oui	Oui	3	Divorcée	Oui	1	<30	Non	Pas d'allaitement	14	Oui	>=50	Non	Non	Non	0
195	25	52	1,58	20,82999519	Non	Non	0	Célibataire	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	9	Non	Non ménopausée	Non	Non	Oui	5
196	36	74	1,6	28,90625	Oui	Non	0	Célibataire	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	12	Non	Non ménopausée	Non	Non	Non	0
197	33	53	1,68	18,77834467	Oui	Oui	1	Mariée	Oui	1	>=30	Non	Pas d'allaitement	14	Non	Non ménopausée	Non	Non	Non	0
198	44	78	1,55	32,46618106	Oui	Oui	1	Célibataire	Oui	2	>=30	Oui	Moins d'un an	12	Oui	<50	Non	Non	Non	0
199	55	95	1,75	31,02040816	Oui	Non	0	Mariée	Oui	2	<30	Oui	Moins d'un an	14,5	Oui	>=50	Non	Non	Non	0
200	45	63,5	1,65	23,3241506	Oui	Oui	2	Mariée	Oui	2	<30	Non	Pas d'allaitement	14	Non	Non ménopausée	Non	Non	Non	0
201	44	60	1,56	24,65483235	Oui	Non	0	Mariée	Oui	2	<30	Non	Pas d'allaitement	11	Non	Non ménopausée	Oui	Non	Oui	4
202	46	66	1,5	29,33333333	Oui	Non	0	Mariée	Oui	3	<30	Non	Pas d'allaitement	15	Non	Non ménopausée	Non	Non	Non	0
203	48	72	1,6	28,125	Oui	Oui	1	Mariée	Oui	3	<30	Non	Pas d'allaitement	13	Non	Non ménopausée	Non	Non	Non	0

204	55	59	1,59	23,33768443	Oui	Non	0	Mariée	Oui	4	<30	Oui	Moins d'un an	12	Oui	>=50	Non	Non	Oui	3
205	67	86	1,6	33,59375	Oui	Oui	1	Mariée	Oui	2	>=30	Oui	Moins d'un an	15	Oui	<50	Non	Non	Oui	7
206	50	67	1,65	24,6097337	Non	Oui	1	Divorcée	Oui	3	<30	Non	Pas d'allaitement	11	Non	Non ménopausée	Non	Non	Oui	1
207	46	76	1,59	30,06210197	Oui	Non	0	Mariée	Oui	3	<30	Oui	Plus d'un an	11	Non	Non ménopausée	Non	Oui	Oui	2
208	37	85	1,73	28,40054796	Non	Non	0	Mariée	Oui	2	>=30	Oui	Moins d'un an	12	Non	Non ménopausée	Non	Non	Non	0
209	54	55	1,74	18,16620425	Oui	Oui	1	Mariée	Oui	2	<30	Oui	Moins d'un an	14	Oui	>=50	Non	Non	Oui	5
210	52	80	1,75	26,12244898	Oui	Oui	2	Mariée	Oui	2	<30	Non	Pas d'allaitement	13	Non	Non ménopausée	Oui	Non	Non	0
211	35	65	1,68	23,03004535	Non	Non	0	Mariée	Oui	1	<30	Oui	Plus d'un an	13	Non	Non ménopausée	Non	Non	Non	0
212	55	70	1,65	25,71166208	Oui	Oui	1	Mariée	Oui	2	>=30	Oui	Moins d'un an	11	Oui	>=50	Non	Oui	Non	0
213	38	67	1,69	23,45856238	Non	Non	0	Mariée	Oui	1	<30	Oui	Plus d'un an	13	Non	Non ménopausée	Non	Non	Non	0
214	44	73	1,7	25,25951557	Non	Non	0	Mariée	Oui	3	<30	Oui	Plus d'un an	14	Non	Non ménopausée	Non	Non	Non	0
215	29	59	1,63	22,20633069	Non	Non	0	Célibataire	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	13	Non	Non ménopausée	Non	Non	Oui	2
216	48	75	1,73	25,05930703	Non	Non	0	Mariée	Oui	2	<30	Oui	Plus d'un an	11	Non	Non ménopausée	Non	Non	Non	0

217	55	80	1,67	28,68514468	Oui	Oui	2	Mariée	Oui	2	>=30	Oui	Moins d'un an	10	Oui	>=50	Oui	Non	Oui	1
218	27	65	1,71	22,22906193	Non	Non	0	Célibataire	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	13	Non	Non ménopausée	Non	Non	Non	0
219	43	80	1,74	26,42356982	Oui	Oui	1	Mariée	Oui	1	>=30	Non	Pas d'allaitement	12	Non	Non ménopausée	Non	Oui	Oui	2
220	23	55	1,68	19,48696145	Non	Non	0	Célibataire	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	13	Non	Non ménopausée	Non	Non	Oui	1
221	28	62,5	1,69	21,88298729	Non	Non	0	Mariée	Oui	1	<30	Oui	Moins d'un an	12	Non	Non ménopausée	Non	Non	Non	0
222	50	80	1,75	26,12244898	Non	Non	0	Mariée	Oui	2	>=30	Oui	Moins d'un an	14	Non	Non ménopausée	Non	Non	Non	0
223	55	76	1,8	23,45679012	Non	Oui	1	Mariée	Oui	3	<30	Oui	Plus d'un an	11	Oui	>=50	Non	Non	Non	0
224	37	69	1,77	22,02432251	Non	Non	0	Mariée	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	12	Non	Non ménopausée	Non	Non	Non	0
225	49	72	1,74	23,78121284	Oui	Oui	2	Mariée	Oui	2	<30	Oui	Plus d'un an	12	Non	Non ménopausée	Non	Non	Oui	2
226	55	85	1,68	30,11621315	Oui	Non	0	Veuve	Oui	2	>=30	Oui	Moins d'un an	10,5	Oui	<50	Non	Non	Non	0
227	26	76	1,7	26,29757785	Non	Non	0	Célibataire	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	13,5	Non	Non ménopausée	Non	Non	Non	0
228	39	80	1,8	24,69135802	Non	Oui	1	Mariée	Oui	2	>=30	Oui	Moins d'un an	11	Non	Non ménopausée	Non	Non	Non	0
229	60	77	1,67	27,60945176	Oui	Oui	2	Veuve	Oui	3	<30	Oui	Moins d'un an	10	Oui	<50	Non	Non	Non	0
230	56,5	95	1,83	28,36752366	Non	Non	0	Mariée	Oui	3	<30	Oui	Plus d'un an	13	Oui	>=50	Non	Non	Oui	1
231	55	78	1,75	25,46938776	Oui	Non	0	Mariée	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	12	Oui	>=50	Non	Non	Non	0
232	40	119	1,8	36,72839506	Non	Non	0	Mariée	Oui	2	<30	Oui	Moins d'un an	11	Non	Non ménopausée	Non	Non	Non	0
233	37	67	1,68	23,73866213	Non	Oui	1	Célibataire	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	13	Non	Non ménopausée	Non	Non	Non	0
234	45	80	1,56	32,8731098	Oui	Oui	2	Mariée	Oui	2	<30	Oui	Moins d'un an	11	Non	Non ménopausée	Non	Non	Non	0
235	48	76,5	1,75	24,97959184	Non	Non	0	Mariée	Oui	2	<30	Oui	Moins d'un an	12	Non	Non ménopausée	Non	Non	Non	0
236	60	86	1,8	26,54320988	Oui	Non	0	Veuve	Oui	3	<30	Oui	Moins d'un an	12	Oui	>=50	Oui	Non	Non	0
237	57	77	1,64	28,62879239	Non	Oui	2	Mariée	Oui	4	<30	Oui	Plus d'un an	14	Oui	<50	Non	Non	Non	0

238	43	70	1,6	27,34375	Non	Oui	1	Mariée	Oui	1	<30	Oui	Plus d'un an	11,5	Non	Non ménopausée	Non	Non	Non	0
239	25	68	1,7	23,52941176	Non	Non	0	Célibataire	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	10	Non	Non ménopausée	Non	Non	Oui	2
240	43	76,5	1,59	30,25987896	Non	Non	0	Mariée	Oui	2	<30	Oui	Plus d'un an	13	Non	Non ménopausée	Non	Non	Oui	1

241	38	75	1,79	23,40750913	Non	Non	0	Célibataire	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	12	Non	Non ménopausée	Non	Oui	Non	0
242	46	82	1,65	30,11937557	Oui	Oui	1	Mariée	Oui	2	<30	Oui	Moins d'un an	12,5	Non	Non ménopausée	Non	Non	Non	0
243	34	70	1,56	28,76397107	Non	Non	0	Mariée	Oui	1	>=30	Oui	Moins d'un an	11	Non	Non ménopausée	Non	Non	Non	0
244	56	86	1,74	28,40533756	Oui	Oui	2	Mariée	Oui	2	<30	Oui	Moins d'un an	13	Oui	>=50	Oui	Non	Non	0
245	45	77	1,77	24,57786715	Non	Non	0	Mariée	Oui	1	>=30	Oui	Plus d'un an	11	Non	Non ménopausée	Non	Non	Oui	3
246	58	79	1,73	26,3958034	Non	Non	0	Veuve	Oui	4	<30	Oui	Moins d'un an	9	Oui	>=50	Non	Non	Non	0
247	26	76	1,65	27,91551882	Non	Oui	1	Célibataire	Non	0	Pas encore	Pas d'enfants	Pas d'allaitement	11	Non	Non ménopausée	Non	Non	Non	0

Avec :

X_1 : L'âge ; X_2 : Le poids (Kg) ; X_3 : La taille (m) ; X_4 : L'IMC (kg/m²) ; Y : Le cancer du sein ; X_5 : Les antécédents familiaux ; X_6 : Nombre de membre de famille malade ; X_7 : Situation matrimoniale ; X_8 : Parité ; X_9 : Nombre d'enfants ; X_{10} : Age à la première naissance ; X_{11} : Allaitement au sein ; X_{12} : Durée d'allaitement ; X_{13} : La ménarche ; X_{14} : Ménopausée ; X_{15} : Age à la ménopause ; X_{16} : Traitement hormonaux de la ménopause ; X_{17} : Pilules contraceptives ; X_{18} : Activité physique ; X_{19} : Nombre de pratique du sport (Par semaine).

Annexe 3 : description des attributs de la base de données.

Les attributs	Le type d'attribut	Les valeurs d'attributs (Modalités)
Cancer_DuSein	Booléen	0 : Non 1 : Oui
Age	Continue	/
IMC	Continue	/
Antecedent_Fami	Booléen	0 : Non 1 : Oui
Nmbre_deMemFam_Malades	Discret	/
Situation_Matrimoniale	Qualitatif	0 : Mariée 1 : Célibataire 2 : Veuve 3 : Divorcée
Parite	Booléen	0 : Non 1 : Oui
Nombre_d'enfants	Discret	/
Age_aLaPrem_Naiss	Qualitatif	- Pas encore - <30 - >=30
Allaitement_auSein	Qualitatif	- Pas d'enfants - Non - Oui
Duree_Allait	Qualitatif	0 : Pas d'allaitement 1 : Moins d'un an 2 : Plus d'un an
La_Menarche	Continue	/
Menopausee	Booléen	0 : Non 1 : Oui
Age_Menopause	Qualitatif	0 : Non ménopausée 1 : <50 2 : >=50
THM	Booléen	0 : Non 1 : Oui
Pills_contra	Booléen	0 : Non 1 : Oui
Activite_phys	Booléen	0 : Non 1 : Oui
Nombre_dePrat	Discret	/

Table des matières