République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Saad Dahlab Blida
USDB

Faculté des sciences

Département d'informatique

Mémoire de fin d'études pour l'obtention du diplôme de master en informatique

Option : Ingénierie de logiciel

Thème :

Développement d'un outil d'analyse prédictive :

Application dans le domaine médical.

Réalisé par :

M^lle Bourouis *Zakia Nedjla*
M^lle Elmouiah *Mounia*

Encadré par :

M^me H. ABED

Membres des jury :
Ferfera Sofiane **(Président)**
Hadj Yahia Ouahid
Chikhi Nacim Fateh

Promotion : 2014 / 2015

« L'homme est sage tant qu'il cherche la sagesse,
Mais dès qu'il croit l'avoir trouvée, il perd la tête. »

(Proverbe arabe)

Remerciements

Nous aimerions en premier lieu remercier ALLAH *l'unique et le seul, qui nous a donné la volonté, le pouvoir et le courage pour la réalisation de ce travail.*

Nous tenons à remercier notre promotrice Mme ABED *de nous avoir confiées ce travail. Nous la remercions énormément pour son soutien, son support, sa clairvoyance, son intérêt et son savoir-faire qui nous ont été d'une aide inestimable.*

Nous remercions également toutes les personnes qui nous ont guidées, aidées et conseillées : Mr AIT AKKACH, Mr EL MOUSSAOUI, Mr HAMMOUDA, Mr DJENNOURI, Mr LAMRANI, ainsi qu'à toute l'équipe du CERIST qui nous a bien accueillies spécialement Mr BERROUK, Mr AMAREDJ.

Que les membres du jury trouvent ici le témoignage de notre reconnaissance pour avoir bien accepté d'évaluer notre modeste travail.

Sans oublier toute personne qui nous a aidées ou tenté de le faire. Merci !

ملخص :

في عام 2015، التحليل التنبؤي لم يعد سحرا ولا خيالا. على العكس تماما، يستوجب على هذه الأداة الإحصائية للتنبؤ والتصنيف أن تكون في قلب الكثير من المجالات بما فيها الطب. المبدأ الأساسي هو استخدام المعلومات من البيانات المتعلقة بمجال الدراسة لتشكيل التوقعات، مما يسمي تعليم الآلة أو التعليم الاصطناعي.

العمل الحالي يسلط الضوء على تطبيق الإنحدار اللوجستي لإستخراج نموذج قادر على حساب إحتمال الحدث المدروس.

تشكيل نموذج الإنحدار اللوجستي يمر بخطوتين أساسيتين : التحليل أحادي المتغير أين يتم إقتناء المتغيرات الواجب إبقاءها في النموذج النهائي عن طريق دراسة تأثير كل متغير على القيمة المتنبأة، و التحليل متعدد المتغيرات أين تتم دراسة الترابط مابين مختلف متغيرات النموذج.

كنهاية لهذه الدراسة، تمت برمجة أداة حساب إحتمال حدوث الحدث.

الكلمات الدلالية : التحليل التنبؤي، تحليل المعلومات، الإنحدار اللوجستي، التعليم الألي، التعليم الإصطناعي

Résumé :

En 2015, l'analyse prédictive ne relève ni de la magie ni de la fiction. Au contraire, cet outil statistique de prévision et de classification se doit d'être au cœur de plusieurs domaines notamment en médecine. Le principe de base est d'utiliser les informations extraites des données relatives au domaine d'étude pour former des prédictions, ce qu'on appelle le machine learning ou l'apprentissage artificiel.

Le travail actuel se focalise sur l'application de la régression logistique pour générer un modèle capable de calculer la probabilité de la survenue de l'évènement étudié. L'élaboration d'un modèle de régression logistique passe par deux étapes majeures : l'analyse univariée où la sélection des variables à retenir dans le modèle final sera effectuée en mesurant l'influence de chacune d'elles sur la variable à prédire, et l'analyse multivariée où on étudie l'association entre toutes les variables du modèle.

La finalité de cette étude est un outil de calcul de probabilité de déclenchement de l'évènement.

Mots clés : Analyse prédictive, analyse des données, régression logistique, apprentissage automatique, machine learning, apprentissage artificiel.

Abstract :

In 2015, predictive analysis is neither magical nor fictional. On the opposite, this statistical tool of prevision and classification must be at the heart of several fields notably in medicine. The basic principle is to use information extracted from study field relative data in order to form predictions, which we call machine learning or artificial learning.

The actual work focuses on the application of logistic regression to generate a model that is capable to calculate the studied event outcome probability. The elaboration of a logistic regression passes by two major steps : the univariate analysis where the selection of variables to be remained in the final model will be done by measuring the influence of each one of them on the predicted variable, and the multivariate analysis where we study the association between all the model's variables.

The finality of this study is a tool to calculate the probability of the event launching.

Key words : predictive analysis, data analysis, logistic regression, machine learning, artificial learning.

SOMMAIRE

LISTE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION

« Le commencement de toutes les sciences, c'est l'étonnement de ce que les choses sont ce qu'elles sont »

Aristote

L'analyse prédictive est un processus analytique polyvalent pouvant être appliqué à des secteurs aussi variés que la vente au détail, dans le but de faire grimper les ventes, les programmes de santé, pour suivre les éclosions de maladies! Par conséquent, il ne s'agit pas d'un processus simple à définir ou à décrire, et ses répercussions pourraient être nulles ou importantes, selon l'utilisation qui en est faite. En outre, il importe de souligner que le concept d'analyse prédictive est étroitement lié à des notions de forage des données déjà connues, mais qu'elle permet d'étendre les inférences au-delà de l'analyse de plusieurs facteurs tels les tendances rétrospectives et d'en arriver à un résultat plus prospectif et anticipatoire.

L'apport de ce type d'approche fait rêver beaucoup de décideurs : pouvoir comprendre et anticiper des évènements avant qu'ils ne surviennent, afin de mettre en place des pistes d'amélioration ou des actions correctrices, peut sembler directement sorti des cerveaux les plus farfelus amateurs de science-fiction. Pourtant, ces méthodes sont parfois anciennes (plusieurs dizaines d'années) et reposent en toute simplicité sur l'analyse des données réelles à disposition.

La façon dont les modèles prédictifs apportent de la valeur ajoutée constitue un concept très simple : ils permettent de prendre des décisions plus éclairées, offrent une cohérence accrue de façon plus rapide et à moindre coût. Ils améliorent les décisions humaines en augmentant leur efficacité et leur efficience. Dans certains cas, ils permettent même d'automatiser un processus complet de prise de décision et les réactions aux traitements médicaux en constitue un exemple concret du fait que les modèles liant les symptômes aux traitements sont de plus en plus répandus. Par exemple, un modèle peut prédire la probabilité qu'un patient présentant certains symptômes soit en fait victime d'une crise cardiaque, aidant ainsi le personnel des

urgences à déterminer le traitement requis et le niveau d'urgence. Le rôle du modèle prédictif permet non seulement de comprendre un phénomène mais surtout de retenir les données pertinentes noyées dans une masse d'information, du coup le bon déroulement de ce genre d'analyse réclame surtout des données, lesquelles ne manquent pas en général.

Au cours des dernières années, l'analyse prédictive est passée de technique avant-gardiste peu répandue à une arme concurrentielle dont la portée se développe rapidement. L'adoption croissante de l'analyse prédictive est alimentée par des tendances convergentes : le phénomène des données massives, l'amélioration des outils d'analyse de données et un afflux constant de réussites démontrées dans le cadre de nouvelles applications. Aujourd'hui, l'analyste moderne dirait sûrement **« Donnez-moi suffisamment de données et je prédirai n'importe quoi. »**

1. Problématique :

Longtemps considérée comme complexe, insurmontable et à la portée de quelques érudits voire impossible à effectuer ; la modélisation prédictive est aujourd'hui, faisable, très simple d'accès et est dans certains cas une pratique fondamentale et indispensable. Les dernières nouveautés en matière de modélisation prédictive accélèrent cette démarche, en brassant un nombre d'information, le tout en quelques secondes. Ces informations auparavant collectées, stockées mais rarement analysées sous l'angle prédictif.

A la confluence des mathématiques et de l'informatique, le Machine Learning est un ensemble de modèles et d'algorithmes permettant à des systèmes d'apprendre automatiquement à partir d'une masse de données et d'effectuer des tâches variées. Les progrès accomplis au cours de la dernière décennie dans cette discipline en plein essor ont conduit à la création d'algorithmes et de modèles de prédiction toujours plus performants. Ils constituent, combinés à la ressource en données et en puissance de calcul, un levier de transformation puissant dans plusieurs domaines.

La problématique ici consiste à concevoir un système médical faisant un calcul de prédiction du cancer de sein en analysant une manne de données. La nécessité d'automatiser le diagnostic médical est devenu indispensable et pousse à aller plus loin dans la recherche des solutions comme nous essayons de faire.

2. Objectifs :

L'objectif plus large assigné à ce travail est la mise en place d'un outil de prédiction en tirant profit des méthodes statistiques pour arriver finalement à prédire un évènement précis en s'appuyant sur un nombre d'entrées. Et pour ce faire, on va :

- Réaliser une collecte de données, les comprendre en étudiant chaque variable et sa nature (transformation si besoin etc.) ;
- Appliquer les principes d'apprentissage artificiel et des statistiques (ce qu'on appelle l'apprentissage statistique) pour avoir la possibilité de créer un modèle prédictif précis sur le futur ;
- Evaluer le modèle élaboré afin d'en vérifier la fiabilité en le testant sur les données existantes et appliquer des prédictions aux nouvelles données ;
- Appliquer le résultat de cette étude pour l'aider à la prise de décision quotidienne et pour obtenir les résultats en automatisant les décisions de son domaine d'application.

CHAPITRE 1 : ETAT DE L'ART

« L'expérience fait l'art, l'inexpérience la fortune. On fait des découvertes en cherchant et des trouvailles par hasard. »

Joseph Joubert

1. Introduction :

L'analyse prédictive englobe une variété de techniques issues des statistiques, dont l'objectif est d'associer une probabilité à un évènement futur. Le calcul de cette probabilité étant fondé sur l'observation du passé et toutes les données passées caractérisant le comportement à prédire.

Historiquement, la Statistique s'est beaucoup développée autour de ce type de problèmes et a proposé des modèles établis en fonction des variables propres au domaine d'application. Il s'agit alors d'estimer les paramètres du modèle à partir des observations. Dans la même situation, la communauté informatique parle plutôt d'apprentissage visant le même objectif, notamment l'apprentissage machine (ou machine learning).

L'objectif général est donc celui de modélisation qui peut se préciser en sous-objectifs à définir clairement préalablement à une étude car ceux-ci conditionnent en grande part les méthodes qui pourront être mises en œuvre.

Des paramètres importants du problème sont les dimensions : n nombre d'observations et p nombre de variables observées sur cet échantillon. Lorsque les méthodes statistiques traditionnelles se trouvent mises en défaut pour de grandes valeurs de p, éventuellement plus grande que n, les méthodes récentes d'apprentissage sont des recours pertinents car efficaces. L'étude des données se focalise donc sur les pratiques de l'apprentissage machine et de la statistique. Les développements méthodologiques ont pris depuis le début du siècle la dénomination d'apprentissage statistique.

2. Les approches numériques de l'aide au diagnostic :

2.1 Les approches statistiques : [1]

Les méthodes statistiques reposent sur des techniques de classification multifactorielle, en particulier l'analyse discriminante. Schématiquement, on considère un ensemble de N observations (des patients), décrites par k variables (les signes, les symptômes et les résultats d'examens complémentaires), reparties en n catégories (les diagnostics).

L'analyse discriminante vise à produire un nouveau système de représentation, constitue des combinaisons linéaires des variables initiales, qui permet de séparer au mieux les catégories. Il s'agit ainsi de construire une fonction de classement permettant de prédire la catégorie d'un individu à partir des valeurs prises par les variables qui le caractérisent. Ainsi, cette technique se rapproche des techniques supervisées en apprentissage automatique (dit aussi artificiel). De nombreuses méthodes ont été proposées. Par exemple, si on cherche à discriminer les sujets malades des non-malades (ici n = 2), on peut dire qu'il s'agit, dans un espace a k dimensions correspondant aux variables décrivant les patients, de trouver le plan qui sépare au mieux les points correspondant aux malades et les points correspondant aux sujets sains. Cette fonction obtenue sur une population d'apprentissage, et testée sur un autre échantillon de données afin d'évaluer sa validité, est calculée sur tout nouveau patient pour lequel on veut poser le diagnostic de la maladie.

2.2 Les approches probabilistes : application du théorème de Bayes : [1]

De la même manière, on dispose d'un ensemble de N observations, chacune étant décrite par k variables X_j formant le vecteur X, et reparties en n catégories, les n diagnostics considérés D_i. Pour tout nouveau patient, le théorème de Bayes permet de calculer les probabilités *a posteriori* des différentes hypothèses diagnostiques D_i :

$$\forall i = 1, \ldots n,\ P(D_i/X) = P(X/D_i).P(D_i) / P(X)$$

Les $P(D_i)$ représentent les probabilités *a priori* des différents diagnostics et sont estimées par les fréquences des D_i calculées sur la base de référence. En supposant que les diagnostics Di sont exhaustifs (l'ensemble des D_i recouvre bien tous les diagnostics possibles) et exclusifs (on ne peut avoir Di et Dj en même temps), on a :

$$\forall i = 1, \ldots n,\ P(D_i/X) = P(X/D_i).P(D_i) / \sum_{i=1}^{n} P(X/D_i).P(D_i)$$

L'approche bayésienne a donné lieu à de nombreuses applications notamment par l'équipe de Dombal à Leeds sur le diagnostic des douleurs aigues de l'abdomen. Chaque patient est défini par une quarantaine de variables, telles que la topographie de la douleur, les facteurs d'exacerbation et d'apaisement, l'existence de nausées, de vomissements, la présence de fièvre, etc. L'hypothèse d'indépendance des signes permet d'écrire :

$$P(X / D_i) = \prod_{j=1}^{k} P(X_j/D_i)$$

Avec huit diagnostics, l'appendicite, la cholécystite aigue, l'occlusion du grêle, la pancréatite, la perforation d'ulcère, la diverticulité aigue, les douleurs abdominales non spécifiques, et « les autres douleurs », le niveau de performance du système informatique est de 91,8 %, significativement supérieur à celui des experts humains du domaine (79,6 %). Pourtant, en dépit de leurs bonnes performances, les systèmes numériques n'ont pas été utilisés en routine clinique. D'une part, les approches numériques, qu'elles soient statistiques ou probabilistes, s'appuient sur une base de données de référence permettant, par exemple dans le cas des approches probabilistes, l'apprentissage des probabilités *a priori* $P(D_i)$ et marginales $P(X/D_i)$, et on a montré que ces bases de référence devaient être locales au site d'utilisation des systèmes probabilistes d'aide au diagnostic, ce qui en limitait la diffusion. Par ailleurs, la pauvreté des interfaces, fondées sur le simple affichage des probabilités des hypothèses diagnostiques, sans aucune explication, n'a pas convaincu les médecins. La figure **(Figure 1.1)** représente un exemple d'interface extrait du système de Leeds.

```
POSSIBLE DIAGNOSES
  APPEND DIVERT PERFDU NONSAP CHOLEC SMBOBT PANCRE
PROBABILITIES ARE
     0·0    0·0     2·7     0·0     0·9     3·1    93·2

CLINICIANS DIAGNOSIS
     PRIMARY    -CHOLEC
     SECONDARY -SMBOBT

COMPUTERS DIAGNOSIS
     PRIMARY    -PANCRE  93·2
     SECONDARY -SMBOBT   3·1

NEITHER OF YOUR DIAGNOSES SEEM LIKELY. PROBABILITIES INDICATE
PANCRE AS PRIME POSSIBILITY

++ SUGGEST CHECKING THE FOLLOWING.....
AMYLASE
TENDERNESS....
SITE PRESENT
```

Figure 1.1 – Affichage des différentes probabilités a posteriori avec commentaires et recommandations.

2.3 Le cas particulier des scores cliniques : [1]

De très nombreux scores cliniques ont été construits et valides pour aider le médecin dans sa démarche diagnostique. Certains sont connus de la plupart des médecins comme le Mini Mental Score (MMS) qui est un test de référence dans le dépistage des démences, ou le test de Fagerstrom qui permet d'évaluer la dépendance tabagique.

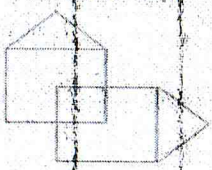
Le principe de ces scores diagnostiques est le suivant : le médecin pose un nombre fixe de questions standardisées et enregistre la réponse qui peut être binaire (vrai/faux), ordinale ou numérique. Le score final prend souvent la forme d'une somme, pondérée ou non, des résultats des réponses aux questions. Néanmoins, certains scores peuvent avoir des formes plus complexes.

Des seuils permettent de conclure (par exemple : détérioration intellectuelle absente, légère, importante). Dans le cas du MMS, le médecin pose au patient trente questions standardisées et compte les réponses justes. Plusieurs capacités cognitives sont successivement explorées :

- l'orientation dans le temps et dans l'espace ;
- l'apprentissage et la transcription des informations ;
- l'attention et le calcul mental ;
- le rappel d'informations et la mémorisation ;
- le langage ;
- la capacité d'organiser une série de mouvements dans un but précis (praxie constructive).

L'implémentation informatique du MMS le rend beaucoup plus facile à utiliser en consultation. Le médecin n'a plus qu'à cliquer sur les cases qui conviennent (**Figure 1.2**). Le score total est automatiquement calculé. Pour être utilisables, ces calculs de score doivent être inclus dans les logiciels métiers des médecins de manière à ne pas interrompre leur «workflow».

LANGAGE

Monter et demander le nom : stylo et montre (1 point par item)	0
Faire répéter : "il n'y a pas de mais ni de si ni de et" : 1 point ou 0	0
Faire exécuter un ordre triple : prenez cette feuille de papier, pliez-la et jetez-la par terre (1 point par item correct)	0
Faire lire et exécuter un ordre écrit : "fermez les yeux" : 1 point ou 0	0
Ecriture spontanée : une phrase. Ne pas donner d'exemple (1 point pour une phrase simple. Orthographe et grammaire indifférentes)	0
Faire copier le dessin suivant : 1 point si les deux polygones sont corrects et entrecoupés au niveau de leur angle droit (NB : ce test est très sensible aux atteintes organiques débutantes) 	0
SCORE TOTAL (maximum 30)	0

Le score MONDE est noté à part.

Compter de 2 en 2 ou de 5 en 5 est parfois préférable pour des patients fâchés avec les chiffres.

RESULTATS:

Un score total de 30 permet de rassurer le patient.

Entre 20 et 30, le diagnostic ne peut être posé. Le patient sera revu après un traitement d'épreuve (voir maladie d'Alzheimer) et/ou un laps de temps.

Au dessous de 20, il existe un réel trouble à suivre de près et à explorer

Figure 1.2 - Extrait du Mini Mental Score

3. L'apprentissage machine :

L'apprentissage est toute méthode permettant de construire un modèle de la réalité à partir des données, soit en améliorant partiellement le modèle, soit en créant un nouveau modèle. Il existe deux tendances principales en apprentissage, celle issue de l'intelligence artificielle et qualifiée de symbolique, et celle issue des statistiques et qualifiée de numérique.

L'apprentissage machine ou automatique est l'ensemble des processus permettant à un ordinateur d'accroitre ses connaissances et de modifier son comportement à la suite de ses

expériences et de ses actes passés, il permet d'extraire et d'exploiter automatiquement l'information présente dans un jeu de données. En cela il couvre un vaste champ d'objectifs comme la sélection de variables, la discrimination, la régression, la sélection de modèle, la génération et l'inférence de règles, etc.

3.1 Intérêt de l'apprentissage artificiel :

L'apprentissage artificiel intervient si on n'a pas assez de connaissances explicites pour obtenir un algorithme qui résout le problème en question ; mais nous avons beaucoup d'exemples de la tâche à accomplir (base de données = exemples). La plupart des programmes possèdent aujourd'hui un module d'apprentissage comme par exemple les programmes de reconnaissance des formes, qui sont fondés sur des algorithmes d'apprentissage. Actuellement, l'apprentissage artificiel permet de réaliser plusieurs tâches différentes :

- Assistance des experts humains dans la prise des décisions complexes (aide aux diagnostics médicaux, analyse des marchés financiers).
- Fouilles d'immenses bases de données hétérogènes (data mining).
- Analyse automatique des photos satellites pour détecter certaines ressources sur la terre, ainsi que l'identification des sous-marins.
- Reconnaissance de formes et de la parole humaine et de l'interpréter.
- Recherche d'information (moteur internet…).
- Détection des fraudes dans le domaine de la télécommunication. Etc…

L'objectif de l'apprentissage automatique est de concevoir des machines capables d'évoluer automatiquement grâce à l'expérience. Pour cela, l'apprentissage automatique regroupe l'ensemble des méthodes permettant à une machine de construire un modèle de la réalité à partir des données, soit en améliorant un modèle partiel ou moins général, soit en créant complètement ce modèle. L'apprentissage automatique est à l'heure actuelle une discipline en plein essor et ses domaines d'application sont très nombreux.

4. L'apprentissage statistique :

L'apprentissage statistique correspond au domaine se consacrant au développement d'algorithmes permettant à une machine d'apprendre à partir d'un ensemble de données, ce qui fait, d'y extraire des concepts et patrons caractérisant ces données. Bien que la motivation originale de ce domaine fût de permettre la mise sur pied de systèmes manifestant une intelligence artificielle, les algorithmes issus de ce domaine sont maintenant répandus dans bien d'autres, tel la bio-informatique. Plus spécifiquement, on peut définir un algorithme

d'apprentissage comme suit : « Un algorithme d'apprentissage est un algorithme prenant en entrée un ensemble de données D et retournant une fonction f ». On désigne par conséquent D comme ensemble d'entrainement ou ensemble d'apprentissage et la fonction f comme modèle. Suite à l'exécution d'un algorithme d'apprentissage, on dira que le modèle a été entrainé sur l'ensemble D [2].

4.1 Apprentissage non supervisé :

L'apprentissage non supervisé ou le clustering vise à trouver une structure cohérente au sein d'un ensemble susceptible d'en faciliter l'interprétation, l'analyse et la représentation. Il est utilisé si les classes et leurs nombres sont inconnus. L'apprentissage se ramène dans ce cas à identifier des groupes tels que les exemples les plus similaires appartiennent au même groupe, et que les exemples les plus différents soient séparés dans différents groupes [3]. Autrement dit, dans ce type d'apprentissage, les classes d'appartenance ne sont pas connues à priori et dans certains cas, on ne connait même pas leur nombre.

4.2 Apprentissage supervisé :

L'apprentissage supervisé a un rôle prédictif. Il permet d'évaluer la distribution d'une quantité (eg. la taille d'un individu) sans la mesurer directement, mais en se basant sur des valeurs qui lui sont liées (eg. le poids de la personne) [4].

Typiquement, un tel ensemble est récolté en fournissant l'ensemble des entrées à un groupe de personnes et en leur demandant d'associer à chacune de ces entrées une cible appropriée dans le contexte du problème à résoudre. La tâche d'un algorithme d'apprentissage est alors d'entraîner un modèle qui puisse limiter ce processus d'étiquetage par un humain, c'est-à-dire qui puisse prédire pour une entrée x quelconque la valeur de la cible y qui aurait normalement été donnée par un humain. Cependant, les algorithmes d'apprentissage ne se limitent pas à la modélisation du comportement de l'humain et peuvent être utilisés pour modéliser la relation liant des paires d'entrées et de cibles provenant d'un autre phénomène (e.g la relation entre une action et son prix à la bourse telle que générée par les marches financiers). La nature de l'ensemble Y d'où proviennent les cibles dépendra du type de problème à résoudre [2].

L'apprentissage supervisé vise toujours de construire une fonction ou concept sous-jacent f à partir d'un ensemble d'attributs x_i que l'on nomme ici attributs prédictifs. Selon la nature

de *Y*, nous distinguons généralement deux familles d'apprentissage supervisé : La régression et la classification.

4.2.1 Régression :

Dans les problèmes de régression, l'entrée n'est pas associée à une classe, mais dans le cas général, à une ou plusieurs valeurs réelles (un vecteur). Par exemple, pour une expérience de biochimie, on pourrait vouloir prédire le taux de réaction d'un organisme en fonction des taux de différentes substances qui lui sont administrées. Il existe deux types de régression : logistique et linéaire.

En général, la régression logistique est utile lorsqu'on souhaite être capable de prévoir la présence ou l'absence d'une caractéristique ou d'un résultat en fonction de certaines valeurs ou d'un groupe de variables prédites. Elle est similaire à la régression linéaire mais elle convient aux modèles dans lesquelles les variables sont dichotomiques (qui ne prennent que deux valeurs, 0 et 1 par exemple).

4.2.2 Classification :

Dans les problèmes de classification, l'entrée correspond à une instance d'une classe, et la sortie qui y est associée indique la classe. Par exemple pour un problème de reconnaissance de visage, l'entrée serait l'image bitmap d'une personne telle que fournie par une caméra, et la sortie indiquerait de quelle personne il s'agit (parmi l'ensemble de personnes que l'on souhaite voir le système reconnaître).

5. L'apprentissage automatique et les statistiques classiques :

Du point de vue du problème de l'apprentissage, les statistiques classiques se divisent en deux branches :

5.1 Les statistiques paramétriques :

Dont le cadre suppose que l'on connaît la forme du vrai modèle qui a généré les données, ignorant seulement ses paramètres, et où il s'agit d'estimer au mieux les paramètres du dit modèle à partir d'un échantillon de données fini. C'est le cas par exemple des classifications bayésiennes ou de la régression.

5.2 Les statistiques non paramétriques :

Là, la plupart des études statistiques s'intéressent aux propriétés de convergence et consistance de l'estimateur quand le nombre d'exemples tend vers l'infini.

Les recherches en apprentissage automatique se sont quant à elles concentrées davantage sur des problèmes réels complexes, où il serait absurde de croire que l'on puisse disposer du vrai modèle, et où l'on est également loin d'avoir une quantité illimitée de données. Bien que les statistiques classiques se soient un peu intéressées à ces questions, depuis l'avènement de l'informatique ce champ d'investigation a surtout été exploré par la communauté de l'apprentissage automatique. Par ses origines dans des domaines moins frappés de rigueur et de formalisme mathématique (la neurobiologie et l'électronique/informatique), les recherches en intelligence artificielle symbolique ont pris un chemin davantage empirique, se satisfaisant très bien de produire des modèles mathématiques comme les réseaux de neurones, du moment qu'ils fonctionnaient et donnaient de bons résultats. Dans la mesure où les modèles utilisés étaient plus complexes, les questions de sélection de modèle et du contrôle de leur capacité se sont imposées naturellement avec force. Mais on voit que, bien plus qu'une différence de fond entre les deux domaines, ce qui les sépare est une différence de culture et d'emphase : les études statistiques classiques se sont souvent autolimitées à des modèles se prêtant bien à une analyse mathématique (modèles assez simples, en faible dimension). En comparaison, la recherche en intelligence artificielle était résolument engagée sur la voie de la complexité, avec pour seule limite la capacité du matériel informatique, et poussée par le besoin de mettre au point des systèmes répondant aux problèmes concrets du moment. Néanmoins, avec le temps, le domaine de l'apprentissage automatique a mûri, s'est formalisé, théorisé, et s'est ainsi inéluctablement rapproché des statistiques, au point d'être rebaptisé apprentissage statistique.

6. Modélisation du problème :

Dans la plupart des problèmes, les données observées peuvent être décrites par des valeurs numériques ou symboliques. Soit Y une variable binaire (dichotomique) à expliquer et $X = (X_1, X_2, \ldots, X_p) \in R^p$ p variables explicatives. Par exemple, en épidémiologie, ce modèle est utilisé pour étudier les relations entre une maladie Y et des facteurs de risque X_i.

On a opté pour cette modélisation car la valeur à laquelle on s'intéresse peut être représenté par une variable à deux modalités (appartenance ou non appartenance à une classe). On notera donc les deux modalités de Y par 1 en cas d'appartenance et 0 sinon.

7. Apprentissage versus modélisation : [10]

Une confusion survient souvent entre l'apprentissage et la modélisation, alors que ces deux notions bien que synonymes, impliquent un but différent. L'objectif d'une modélisation est d'expliquer, d'interpréter, d'approcher la réalité. Le choix du modèle est guidé par les critères d'ajustement et l'interprétation du rôle de chaque variable explicative est prépondérante. Tandis que pour l'apprentissage au sens théorique de Vapnik [Vapnik , 1998], les choix sont basés sur des critères de qualité de prédiction (nombre de paramètres, complexité …), car le meilleur modèle n'est pas forcément celui qui approche le mieux la réalité. En effet, la plupart du temps, plus le modèle est complexe, plus il est capable de s'adapter aux données, et moins il est capable de généraliser.

8. Génératif versus discriminatif : [10]

Les modèles génératifs essayent d'estimer la probabilité d'apparition de chaque classe pour un certain nouvel élément. Les modèles discriminatifs essayent d'estimer directement quelle est la classe d'appartenance du nouvel élément. D'une façon générale, plus on a d'exemples, plus il semble intéressant de travailler dans un contexte de classification. Il existe de nombreuses méthodes de classification. Il n'y a pas de méthodes globalement meilleures que les autres, donc une bonne connaissance du problème est nécessaire pour choisir la bonne méthode à utiliser. Le choix de la méthode dépend notamment du problème posé, de la nature des données, des propriétés de la fonction à estimer… De plus, la difficulté intrinsèque du problème dépend la qualité des données. En effet, dans la pratique, les données peuvent être fausses, incomplètes, manquantes, non-exhaustives, les résultats sont donc souvent imprécis. Avec des algorithmes exacts sur des données réelles, les résultats fournis sont justes par rapport aux données, mais pas nécessairement par rapport à la réalité. Avec des algorithmes de la classe des heuristiques, on ne sait pas si les résultats obtenus sont justes, mais en général ils sont cependant satisfaisants.

9. Quelques lois de probabilité :

a) Loi de Bernoulli b(p) : [11]

C'est la loi d'une variable aléatoire X qui ne peut prendre que deux valeurs, 1 avec la probabilité p et 0 avec la probabilité 1-p notée q :

$P(X=1) = p$; $P(X=0) = 1- p = q$; $EX = p$; $Var(X) = pq$.

Tel que : $p \in [0, 1]$.

b) Loi binomiale : [12]

La variable binomiale, S_n, représente le nombre de succès obtenus lors de la répétition de n épreuves *identiques et indépendantes*, chaque épreuve ne pouvant donner que deux résultats possibles.

Ainsi la loi de probabilité suivie par la somme de n variables de Bernoulli où la probabilité associée au succès est p, est la loi binomiale de paramètres n et p.

$$S_n : \Omega^n \rightarrow IR^n$$

$$S_n = \sum_{i=1}^{n} X_i \rightarrow \beta(n,p)$$

c) Loi normale : [13]

La loi normale est l'une des lois de probabilité les plus adaptées pour modéliser des phénomènes naturels issus de plusieurs événements aléatoires. La distribution normale, appelée aussi gaussienne est une distribution continue qui dépend de 2 paramètres μ et σ.

Sa densité est donnée par :

$$f(x) = 1 / (\sigma\sqrt{2\pi}) * e^{-1/2 * ((x-\mu)/\sigma)^2}$$

Où μ représente on espérance, et σ est son écart type.

10. Les variables : [5]

Une variable est une caractéristique dont on peut observer des valeurs différentes au sein d'un groupe de sujets. Une variable peut être de nature **qualitative** ou de nature **quantitative.**

10.1 Variable catégorielle ou qualitative :

Une variable dite **catégorielle** ou **qualitative** est une caractéristique ayant un certain nombre de catégories ou modalités. Quand il s'agit de classer les sujets selon deux catégories, la variable catégorielle est dite **dichotomique** (ou **binaire**).

Par exemple, si l'on dénombre les hommes et les femmes dans un groupe, la variable « sexe » est une variable catégorielle à deux catégories : « hommes » et « femmes ». On peut

également classer les sujets selon qu'ils sont fumeurs ou non-fumeurs, selon qu'ils sont atteints ou non d'une maladie.

Certaines de ces variables catégorielles sont dites **nominales** : chaque classe désigne une catégorie de sujets (elle les nomme). Il n'existe pas d'ordre naturel entre les catégories.
C'est, par exemple, le cas du groupe sanguin : A / B / AB / O ou encore de la situation familiale : marié / vivant en couple / célibataire / divorcé / séparé / veuf.

Pour d'autres variables, il existe un ordre naturel entre les différentes catégories. Ces variables sont dites **ordinales**.
Par exemple, lorsque l'on interroge des sujets sur la sévérité d'une douleur : au lieu de deux catégories (douleur / pas de douleur), on peut classer les individus selon les catégories suivantes : aucune / minime / modérée / sévère / insupportable.

10.2 Variable quantitative :

Les valeurs d'une variable **quantitative** sont obtenues par un instrument de mesure ou le résultat d'un dénombrement. Elles sont souvent accompagnées d'une unité de mesure. Avec une telle variable, on peut toujours répondre à une question commençant par : «combien ...? ?». L'âge, le poids, la pression artérielle systolique et la quantité de sucre dans le sang en sont des exemples.

Le modèle de régression logistique permet d'estimer la force de l'association entre une variable qualitative à deux classes (dichotomique) appelée variable *dépendante* et des variables qui peuvent être qualitatives ou quantitatives appelées variables *explicatives* ou indépendantes. La variable dépendante est la survenue ou non de l'événement étudié et les variables explicatives sont des facteurs susceptibles d'influencer la survenue de l'événement (facteurs d'exposition ou facteurs de confusion).

10.2.1 Transformation de variables quantitative en classes :

La pratique qui reste la plus courante en épidémiologie pour inclure une variable quantitative X dans un modèle de régression est de la transformer en variable qualitative en faisant des classes [6]. Cette nouvelle variable qualitative est de plus souvent analysée comme une variable nominale en la remplaçant par des variables indicatrices des classes construites, sans tenir compte de leur ordre, de sorte que la notion même de variable quantitative est quasiment perdue. Ainsi, la présentation et l'interprétation des résultats sont plus simples et/ou plus adaptées aux besoins. En outre, les classes constituées sont souvent les catégories

utilisées de façon habituelle (par exemple, l'âge en classes de 5 ans) et facilitent donc les comparaisons et la discussion des résultats.

11. Choix de la méthode :

Les deux principaux aspects du problème de prédiction, la classification supervisée et la régression, sont des thèmes usuels de la statistique et de l'apprentissage automatique. Dans notre étude, il s'agit d'élaborer un modèle prédictif en utilisant la régression logistique et non pas une régression linéaire par le fait que la variable à expliquer est qualitative. Du coup, nous nous plaçons dans un contexte de classification binaire car il existe seulement deux groupes à discriminer : arrivée ou non d'un évènement précis.

12. Régression logistique :

La régression est une méthode incontournable en traitement des données en particulier dans une démarche de modélisation. Elle consiste à mettre en relation une variable à expliquer Y avec une ou plusieurs variables explicatives appelées prédicteurs [7].

Lorsque la variable Y ne prend que deux valeurs qui signifient l'appartenance à une catégorie ou à un groupe d'individus, la méthode statistique adaptée est la régression logistique binaire.

12.1 La régression logistique et l'épidémiologie [15] :

La régression logistique propose une manière de modéliser la relation entre une variable qualitative à deux classes à des variables X_i, (i = 1, 2, …, k), qui peuvent être quantitatives ou qualitatives. Ce modèle est utilisé en épidémiologie pour étudier les relations entre une maladie M et des facteurs de risque X_i (i = 1, 2, …, k). Par ce modèle, il est possible d'exprimer la probabilité d'être malade connaissant les valeurs prises par les variables X_i. Si l'on suppose que la variable M a deux états que l'on dénote par 0 et 1 (l'état 1 indiquant la présence de la maladie et l'état 0 son absence), on peut écrire :

$$P(M = 1/X_1, X_2, \ldots, X_k) = 1 / (1 + exp[- (\alpha + \sum_{i=1}^{k} \beta_i X_i)])$$

Où $\alpha, \beta_1, \ldots, \beta_k$ sont les paramètres du modèle.

Cette méthode est utilisable chaque fois que l'état de santé auquel on s'intéresse peut être représenté par une variable à deux modalités (présence ou absence d'un signe, malade ou non malade…).

Le plus souvent, on utilise le modèle logistique en épidémiologie pour mesurer et pour tester l'association entre une maladie M et une exposition E (ou un facteur de risque particulier) en tenant compte de facteurs de confusion.

12.2 Principe de la régression logistique :

En épidémiologie, plusieurs modèles d'analyse multivariée sont couramment utilisés : régression linéaire multiple, régression logistique, régression de Poisson, modèle de Cox, etc. Effectuer une régression, c'est tenter de réduire les données d'un phénomène complexe en une loi mathématique simplificatrice. La fonction logistique (qui a donné son nom au modèle) possède des caractéristiques mathématiques expliquant son emploi dans un modèle d'analyse de données épidémiologiques : elle varie de 0 à 1 comme la probabilité de survenue d'un événement ; sa représentation graphique, de forme sigmoïde, correspond assez fidèlement au modèle de relation entre la survenue d'une maladie et un facteur de risque ; enfin, elle permet le calcul aisé des *odds-ratios* (ou rapports de cotes en français).

Le modèle de régression logistique permet d'estimer la force de l'association entre une variable qualitative à deux classes (dichotomique) appelée variable *dépendante* et des variables qui peuvent être qualitatives ou quantitatives appelées variables *explicatives* ou indépendantes. La variable dépendante est la survenue ou non de l'événement étudié (la maladie dans notre cas) et les variables explicatives sont des facteurs susceptibles d'influencer la survenue de l'événement.

La régression logistique peut être univariée mais son intérêt réside dans son utilisation multivariée puisqu'elle permet, alors, d'estimer la force de l'association entre la variable dépendante et chacune des variables explicatives, tout en tenant compte de l'effet simultané de l'ensemble des autres variables explicatives intégrées dans le modèle. L'association ainsi estimée est dite « ajustée » sur l'ensemble des autres facteurs. Même si des adaptations permettent de l'appliquer à certains cas particuliers, le modèle de régression logistique requiert, en principe, certaines conditions :

1. Indépendance des différentes observations entre elles ;
2. Normalité de la distribution des variables quantitatives intégrées dans le modèle ;
3. linéarité de la relation entre chacune de ces variables quantitatives et la variable dépendante.

12.3 Les étapes de la régression logistique :

La réalisation pratique d'un modèle de régression logistique comporte plusieurs étapes :

Etape 1. Le choix et l'étude des variables explicatives : la qualité d'une régression logistique repose, avant tout, sur cette étape. Ce choix est fondé sur la pertinence clinique et sur la connaissance de facteurs de risque avérés ou supposés. C'est pourquoi, une recherche bibliographique approfondie est, au préalable, obligatoire.

Etape 2. Analyse univariée : on procède dans cette étape à l'analyse des liaisons entre chacune des variables explicatives et la variable dépendante; les *odds-ratios* calculés sont bruts. Deux catégories de variables explicatives pourront être intégrées dans un modèle de départ : celles pour lesquelles l'association avec la variable dépendante est suffisamment forte sans toutefois être trop stricte afin de ne pas omettre d'éventuels facteurs de confusion (p-value inférieure ou égale à 0,20, et non pas 0,05, seuil habituellement retenu) et celles qui ont un intérêt clinique avéré en dehors de tout critère d'association (elles sont rares : ce sont des variables dites « forcées ») [14].

Etape 3. Analyse multivariée : Le but des analyses multivariées est de sélectionner, parmi l'ensemble des liaisons statistiques mises en évidence par les analyses univariées, la ou les covariables qui expliquent de façon indépendante la survie ou la maladie.
Ainsi, l'analyse multivariée décante les variables pour ne retenir au final que celles qui suffisent à expliquer sans redondance le modèle. Elle permet non seulement de désigner les variables indépendantes entre elles et de fournir la force des liaisons sous forme d'odds-ratio, mais aussi d'établir des scores dits prédictifs qui permettent d'approcher un pronostic d'une part, de sélectionner ensuite des populations plus homogènes de malades lors d'études ultérieures [5].

Remarque [5]: On dit souvent que ces variables sont *« indépendantes »* c'est uniquement pour expliquer qu'elles sont indépendantes entre elles. Le terme *« indépendant »* ne signifie donc pas *« indépendant de toute variable possible »* mais seulement *« indépendant des variables étudiées »*. On conçoit qu'il suffit d'inclure dans une analyse une variable dont on veut assurer la promotion en association avec quelques autres dont le caractère lié est assez médiocre pour conférer le statut envié de *« variable indépendante »* à celle qui, en présence de variables mieux choisies, aurait peut-être été éliminée. Cette finesse méthodologique justifie à elle seule la nécessité absolue d'une analyse préliminaire des données acquises par tous les média disponibles de façon à proposer à l'analyse multiple un modèle pertinent.

12.4 Le modèle PROBIT versus LOGIT : [15]

Historiquement, les modèles *logit* ont été introduits comme des approximations de modèles *probit* permettant des calculs plus simples. Dès lors, il n'existe que peu de différences entre ces deux modèles dichotomiques. Ceci s'explique par la proximité des familles de lois logistiques et normales.

Le modèle *probit* correspond à la spécification gaussienne introduite plus haut. Tandis que le modèle *logit* correspond à la loi logistique, introduite spécialement pour ce type de modèle.

12.5 Le modèle logit : [10]

La régression logistique peut être décrite d'une autre manière selon un modèle appelé modèle *logit*. Pour un individu ω, on appelle transformation LOGIT de $\pi(\omega)$ l'expression :

$$\text{Log}\,[\pi(\omega) / (1-\pi(\omega))] = a_0 + a_1X_1 + \ldots + a_jX_j$$

La quantité $\pi/(1-\pi) = P(Y=1/X) / P(Y=0/X)$ exprime un odds c'est-à-dire un rapport de chance. Par exemple, si un individu présente un odds de 2, cela veut dire qu'il a 2 fois plus de chances d'appartenir à la classe 1 qu'à la classe 0.

Posons $C(X) = a_0 + a_1X_1 + \ldots + a_jX_j$, on peut revenir sur π avec la fonction logistique:

$$\pi = e^{C(x)} / (1 + e^{C(x)}) = 1 / (1 + e^{-C(x)})$$

La fonction LOGIT = C(X) est théoriquement définie entre -∞ et +∞. En revanche, $0 \leq \pi \leq 1$ issue de la transformation de C(X) représente une probabilité avec les propriétés inhérentes à une probabilité, entre autres : P (Y=1/X) + P (Y=0/X) = 1.

La figure suivante illustre le tracé de la probabilité conditionnelle P(Y/X) en fonction de la fonction LOGIT (tracé de la fonction logistique).

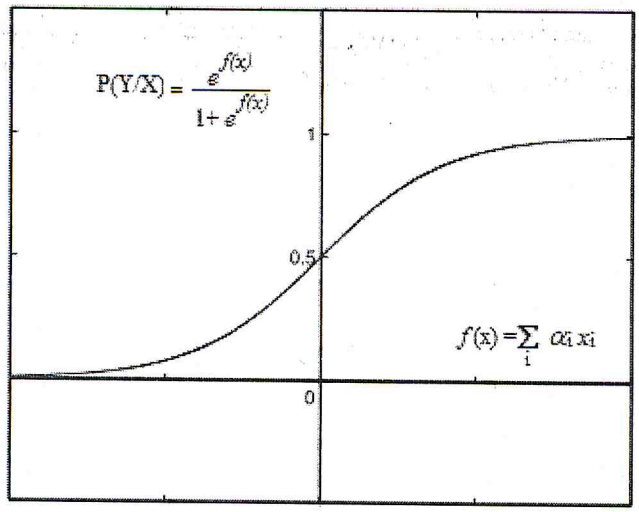


Figure 1. 3 – Tracé de la fonction logistique

12.6 L'estimation des coefficients par maximisation de vraisemblance : [10]

Les variables explicatives X_1, X_2, …, X_j forment une matrice du modèle X dont les lignes sont les vecteurs $(1, x_{i1}, ..., x_{ij})$ et x_{ij} indique la i-ème observation (observation du cas i) de la variable j.

Le modèle $C(x_1, x_2, ..., x_j) = a_0 + a_1 x_1 + \ldots + a_j x_j$ est alors ajusté par la méthode du maximum de vraisemblance. Dans ce but, on résout un système de $(j + 1)$ équations pour les coefficients a_0 et $a_1, \ldots, a_j$, que l'on obtient en annulant les dérivées partielles de la fonction log likelihood $l(a_0, a_1, \ldots, a_p)$.

Pour estimer ces paramètres par la méthode du maximum de vraisemblance, nous devons tout d'abord déterminer la loi de distribution de P(Y/X). Y est une variable binaire définie dans {1, 0}. Pour un individu ω, on modélise la probabilité à l'aide de la loi binomiale B(1, y), avec :

$$P[Y(\omega)/X(\omega)] = \pi(\omega)^{y(\omega)}.(1-\pi(\omega))^{(1-y(\omega))}$$

Cette modélisation est cohérente avec ce qui a été dit précédemment, en effet :

Si $y(\omega) = 1$, alors $P[Y(\omega)=1/X(\omega)] = \pi$

Si $y(\omega) = 0$, alors $P[Y(\omega)=0/X(\omega)] = 1-\pi$

13. Conclusion :

Avec le développement de l'informatique, et du fait de l'augmentation constante des connaissances médicales, l'idée d'utiliser la puissance de calcul et les capacités mémoire des ordinateurs s'est rapidement imposée et de nombreux systèmes informatises d'aide au diagnostic médical ont été développés.

Dans ce chapitre, nous avons essayé de faire un balayage sur tout ce qui concerne l'apprentissage automatique, en définissant ses différentes applications, tâches et typologies, ainsi que sa situation dans les sciences cognitives et l'intelligence artificielle.

Les études sur l'apprentissage artificiel sont en plein essor car c'est un domaine d'actualité qui répond à différents besoins dans la vie quotidienne. Dans le chapitre suivant, nous effectuons la conception de notre nouveau système.

CHAPITRE 2 :
CONCEPTION : APPLICATION DU MODELE LOGISTIQUE POUR L'ANALYSE PREDICTIVE DES DONNEES

« Making a machine that learns is the first step towards making a machine that thinks. »

Dr. Peter J. Bentley,
University College London

1. Introduction :

L'analyse prédictive, parfois appelée analyse avancée, est un terme utilisé pour décrire une série de techniques analytiques et statistiques permettant de prédire des actions ou des comportements futurs. L'analyse prédictive est utilisée pour prendre des décisions au moyen de modèles statistiques afin de tirer des conclusions fiables. Dans ce chapitre, nous présentons les différentes étapes qui mènent à l'obtention de notre instrument de prédiction.

2. Conception du système de prédiction :

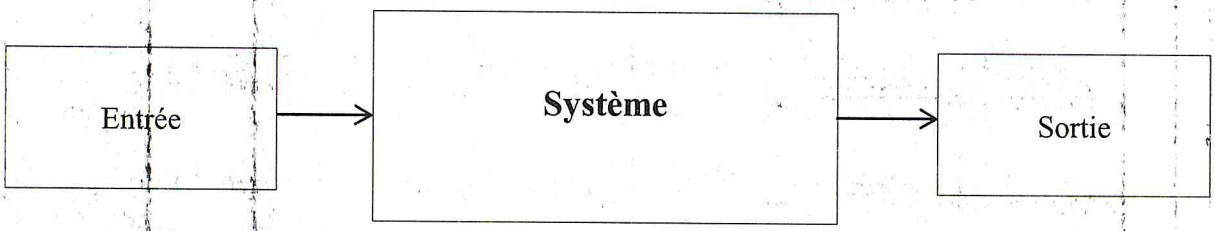


Figure 2. 1 - Schéma de description du système

Le cœur de l'analyse prédictive se fonde sur la capture des relations entre des points de données issues du passé et sur l'utilisation de ces relations pour prédire les résultats futurs. Afin de pouvoir faire des prédictions sur la base de notre ensemble de données (entrées), on utilise plusieurs variables prédictives pour prédire une variable réponse.

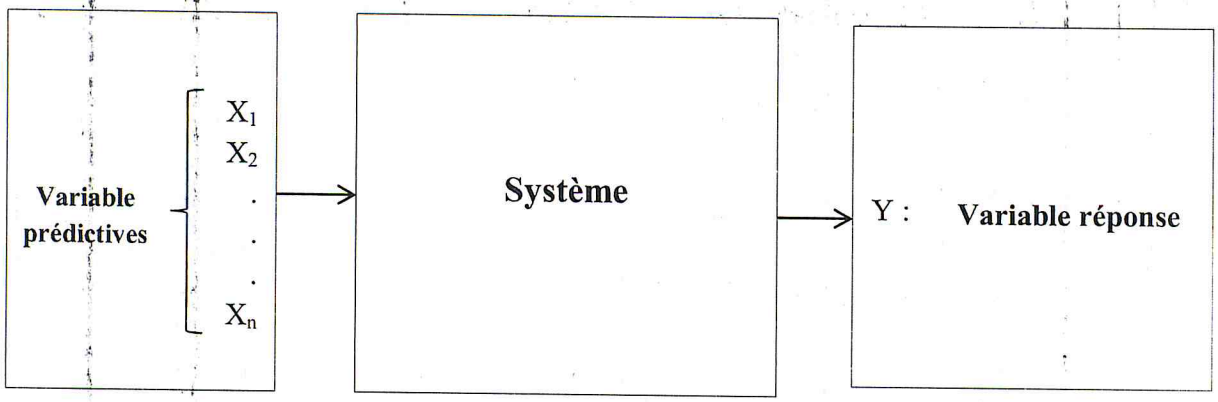


Figure 2. 2 - Schéma de description du processus d'application d'analyse prédictive

Sous sa forme la plus simple, l'analyse prédictive est un support permettant d'établir des prévisions pour une prise de décision. Nous nous intéressons dans la conception de notre système à l'apprentissage supervisé comme catégorie d'analyse prédictive car l'idée est de concevoir un système capable de prédire une variable réponse dont les classes sont connues à priori.

L'apprentissage supervisé est divisé en deux grandes catégories : la régression pour les réponses quantitatives (une valeur numérique), et la classification pour les réponses qui peuvent uniquement prendre quelques valeurs connues, telles que « vrai » ou « faux ».

Régression : la régression est la forme la plus courante de l'analyse prédictive. Elle fait intervenir une variable réponse quantitative (ce qu'on tente de prédire) selon une série de variables prédictives. La relation entre la variable réponse et les prédicteurs de l'ensemble d'apprentissage fournirait un modèle prédictif.

Les méthodes de régression sont nombreuses : régression linéaire multivariée, régression polynomiale, arbres de régression, pour n'en citer que quelques-unes.

Classification : la classification est un autre type d'analyse prédictive communément utilisé. Elle fait intervenir une variable réponse qualitative.

Les types de méthodes de classification sont nombreux : **régression logistique**, arbres de décision, machines à vecteurs de support, forêts aléatoires, k plus proches voisins, la méthode naïve bayésienne, etc.

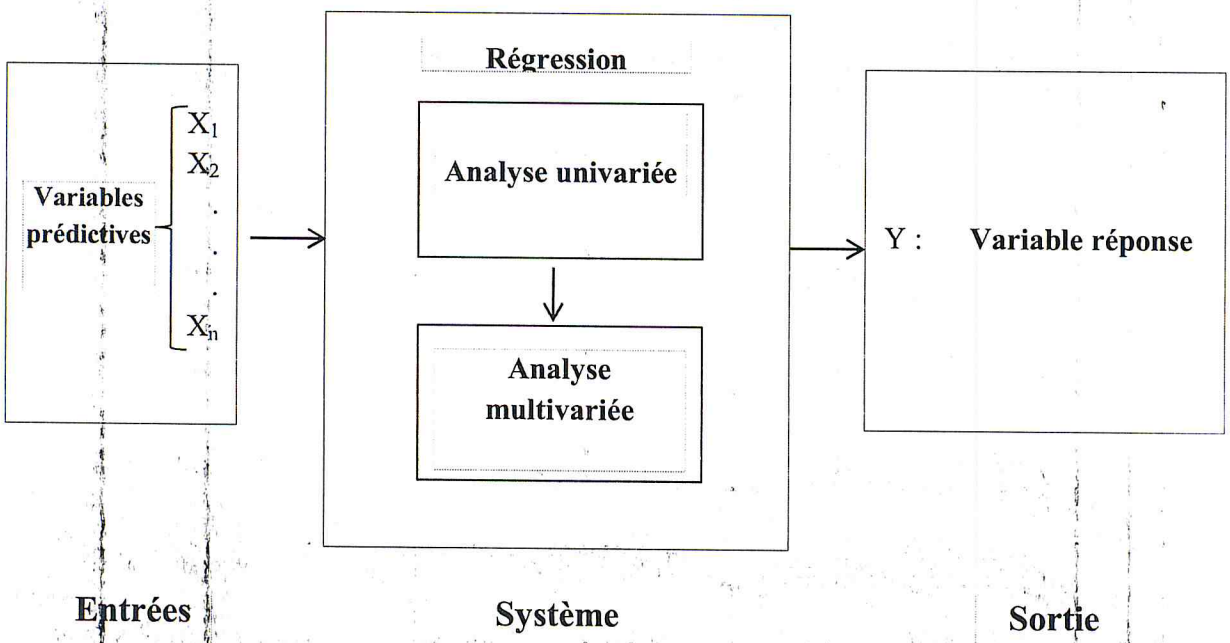


Figure 2. 3 – Schéma de description du processus d'application de la régression logistique et ses deux phases

La régression logistique est une technique statistique connue utilisée pour modéliser les résultats binaires. Utilisation : Exploration et évaluation des facteurs qui contribuent à un résultat comme par exemple rechercher les facteurs qui influencent les clients à se rendre plusieurs fois dans un magasin.

2.1 Formulation :

La tâche que notre système doit effectuer peut être exprimée par la fonction suivante :

$$f : x \longrightarrow u$$

Où X est l'ensemble des objets à classer (aussi appelé espace d'entrée), u est l'ensemble des catégories (aussi appelé espace d'arrivée)

Dans notre démarche, nous nous limiterons à la classification binaire. Dans ce cas, l'ensemble u correspond à {0, +1} La plupart du temps on interprétera +1 et 0 respectivement comme l'appartenance et la non-appartenance à une catégorie déterminée.

2.2 Machine d'apprentissage :

On désigne par machine d'apprentissage, une machine dont la tâche est d'apprendre une fonction au travers d'exemples.

2.3 Ensemble de données :

Le processus d'analyse prédictive est alimenté par l'ensemble de données à utiliser. Pour cela, on divise l'ensemble des données disponibles en deux parties : Un sous ensemble d'apprentissage, dont les données serviront à l'apprentissage (apprendre les paramètres du modèle à proprement dit) et un autre sous-ensemble dit "de validation", qui sert uniquement à évoluer la performance du modèle élaboré. Il est à noter que les variables à prédire de ce sous-ensemble sont aussi déjà classifiées, ce qui permettra de comparer leurs valeurs réelles et celles qu'on a prédite et donc tester la fiabilité de notre modèle.

2.4 Les différentes phases du système :

Jusqu'à présent nous n'avons fait que définir le domaine et l'image de la fonction à estimer. Dans ce qui suit, on parlera des différentes phases de notre système :

a) Phase d'apprentissage :

Dans la phase d'apprentissage, notre système fonctionne de la manière suivante : il examine un ensemble de données dans lequel chaque observation contient des informations sur la variable réponse ainsi que sur les variables prédictives.

Cette tâche de classification se déroulerait de la manière suivante : examiner l'ensemble de données contenant les variables prédictives et *la variable réponse déjà classifiée*. Il apprend ainsi les combinaisons de variables associées à telle ou telle valeur de réponse. C'est cet ensemble de données qu'on appelle l'ensemble d'apprentissage.

Plus formellement, cette phase consiste à sélectionner une fonction $f \in F_\alpha$ c'est-à-dire à trouver une évaluation des paramètres α_i . La sélection de ces paramètres est effectuée par un algorithme d'apprentissage qui reçoit en entrée l'ensemble d'apprentissage. En ce sens, ce sont les données (de l'ensemble d'apprentissage) qui induisent l'apprentissage. L'ensemble des paramètres α_i résultant de l'apprentissage est appelé modèle. Une machine d'apprentissage munie d'un modèle est appelée alors machine entraînée.

b) Phase de prédiction :

Lorsque l'on dispose d'un modèle fiable (validé en utilisant le sous ensemble de validation), on peut faire des prédictions sur de nouveaux exemples. Un outil de prédiction correspond donc à une machine entraînée.

Afin d'exécuter la fonction principale qu'est la prédiction, notre système examinerait les nouvelles observations pour lesquelles aucune information sur la variable à prédire n'est disponible, tout en utilisant le modèle dont les paramètres ont été estimés dans la phase d'apprentissage. Il attribuerait alors des classifications aux nouvelles observations sur la base des classifications dans l'ensemble d'apprentissage.

3. Conclusion :

Nous avons expliqué dans ce chapitre, brièvement, le fonctionnement de notre système en fournissant une vue globale de l'ensemble de ses fonctionnalités sans trop parler de la méthode statistique utilisée et qu'est la régression logistique, cette dernière fera l'objet du prochain chapitre.

CHAPITRE 3 : APPLICATION DU MODELE LOGISTIQUE DANS LE DOMAINE MEDICAL

« Biology and computer science–life and computation–are relater. I am confident that at their great discoveries await those who seek them. »

Leonard Adelman, Scientific American.

1. Introduction :

Les solutions d'analyse prédictive collectent des données et les analysent pour identifier des modèles qui permettront d'évaluer de manière fiable la probabilité d'un résultat ou d'un événement futur. Ceci, en retour, permet une planification, une prise de décision et aide au diagnostic médical plus efficaces et orientées vers l'avenir. Elles fournissent la perspicacité nécessaire afin de prévoir les pathologies susceptibles de se déclencher chez une personne et obtenir les meilleurs résultats possibles éventuellement.
Le majeur but de notre système est l'élaboration d'un instrument de prédiction de la survenue de l'évènement étudié. Sur la base d'un ensemble de données, il devient alors possible de mesurer la probabilité qu'il se produise. L'idée est alors d'imaginer une solution pour prédire les pathologies futures des sujets en fonction de leurs paramètres personnels (historiques etc.). La précision du système à prédire la survenue d'une maladie est mesurée en donnant les facteurs de risque en entrée, puis en appliquant les différentes étapes de l'algorithme de régression logistique. On obtient par la suite un modèle logistique, caractérisant la contribution de ces facteurs à l'apparition de la maladie.

Partie 1 : Contexte médical et généralités

1. Notion de Facteur :

Un facteur est un élément qui influe sur un processus, ou sur un résultat. Il existe trois types de facteur :

2. Facteur de risque : [16]

C'est un facteur associé statistiquement à la survenue d'une maladie ou d'un phénomène de santé. Ils sont encore appelés facteurs favorisants. Ils favorisent l'apparition de la maladie sans en être la cause directe.

2.1 Facteur de protection : [17]

Les facteurs de protection sont les facteurs qui contribuent à réduire la probabilité qu'une personne développe une maladie.

2.2 Facteur de confusion : [18]

On parle de facteur de confusion pour tout facteur lié à la fois à l'affection (tous ses facteurs de risques sont donc concernés) et au facteur étudié. Un facteur de confusion sera lié à un facteur de risque indépendamment de la maladie et ce même facteur de confusion sera lié à une maladie indépendamment du facteur de risque.

3. Les enquêtes épidémiologiques : [19]

Une enquête est une opération qui consiste à rechercher, rassembler, recueillir de l'information, puis à l'analyser en vue de résoudre une ou plusieurs questions spécifiées à l'avance.

Les enquêtes épidémiologiques peuvent concerner l'ensemble de la population : elles sont dites **exhaustives.** Elles peuvent au contraire concerner un échantillon d'effectif réduit, extrait par sondage et représentatif de la population étudiée : il s'agit alors d'enquêtes **par échantillonnage.** Elles se divisent en deux grandes catégories : les **enquêtes expérimentales** et les **enquêtes d'observation**.

3.1 Les enquêtes expérimentales : [19]

Dans les enquêtes expérimentales, l'investigateur contrôle l'attribution aux sujets de l'enquête des facteurs qu'il étudie. L'intérêt est de pouvoir donner une interprétation causale aux associations observées entre exposition et maladie.

3.2 Les enquêtes d'observation : [19]

Ce sont les plus fréquentes en épidémiologie, l'investigateur ne contrôle pas l'exposition aux facteurs d'exposition par opposition aux essais cliniques où l'investigateur contrôle l'allocation du traitement.

3.2.1 Les enquêtes descriptives : [19]

Le principal objectif des enquêtes à visée descriptive est de **mesurer la fréquence d'un problème sanitaire**. Elles sont destinées à compléter le système d'information constitué par les statistiques sanitaires (enregistrement des cas de maladies ou des décès1) et à répondre à des questions ou des hypothèses spécifiques. Ceci implique le choix de populations **représentatives** d'**effectif suffisant** afin d'avoir une vision « exacte » de la réalité.

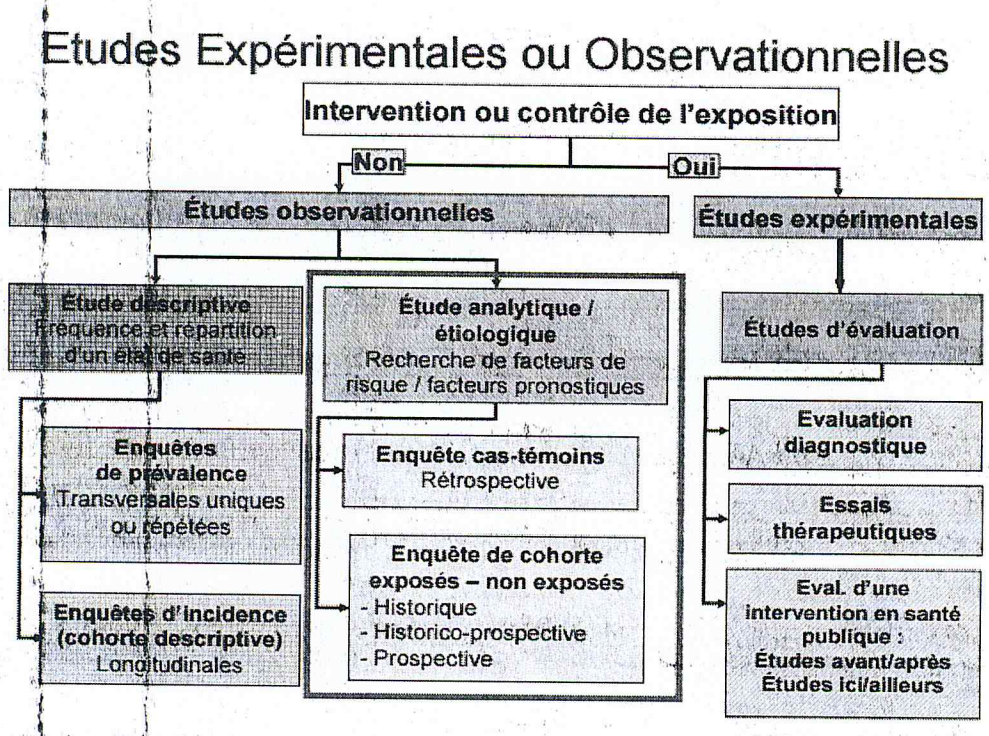


Figure 3. 1 - Les principaux types d'enquêtes en épidémiologie [20]

3.2.2 Les enquêtes analytiques : [20]

Les enquêtes analytiques consistent à étudier les relations existant entre les facteurs de risque et les états pathologiques dans les populations. L'étude de ces associations consiste, à partir d'une observation faite sur un nombre limité de cas, à conclure à **l'existence d'une relation** dont la validité est supposée universelle et à la **quantifier**. Ces études reposent sur un principe simple qui consiste à comparer l'incidence de la maladie chez des sujets exposés et non exposés, ou la fréquence de l'exposition chez des malades et des non malades.

3.2.2.1 Les enquêtes de cohorte ou de type exposé - non exposé : [20]

Les enquêtes de cohorte consistent à comparer la morbidité (ou la mortalité) observée dans un ou plusieurs groupes d'individus initialement indemnes de la maladie et définis en fonction de leur exposition à un facteur de risque soupçonné de la maladie faisant l'objet de l'étude. On les appelle également *enquêtes longitudinales*. Lorsque l'exposition est dichotomique et que l'on compare l'incidence de la maladie d'un groupe exposé à celle d'un groupe non exposé, on parle d'enquête exposé-non exposé. Le terme « cohorte » est utilisé pour désigner le ou les groupes de sujets suivis au cours du temps.

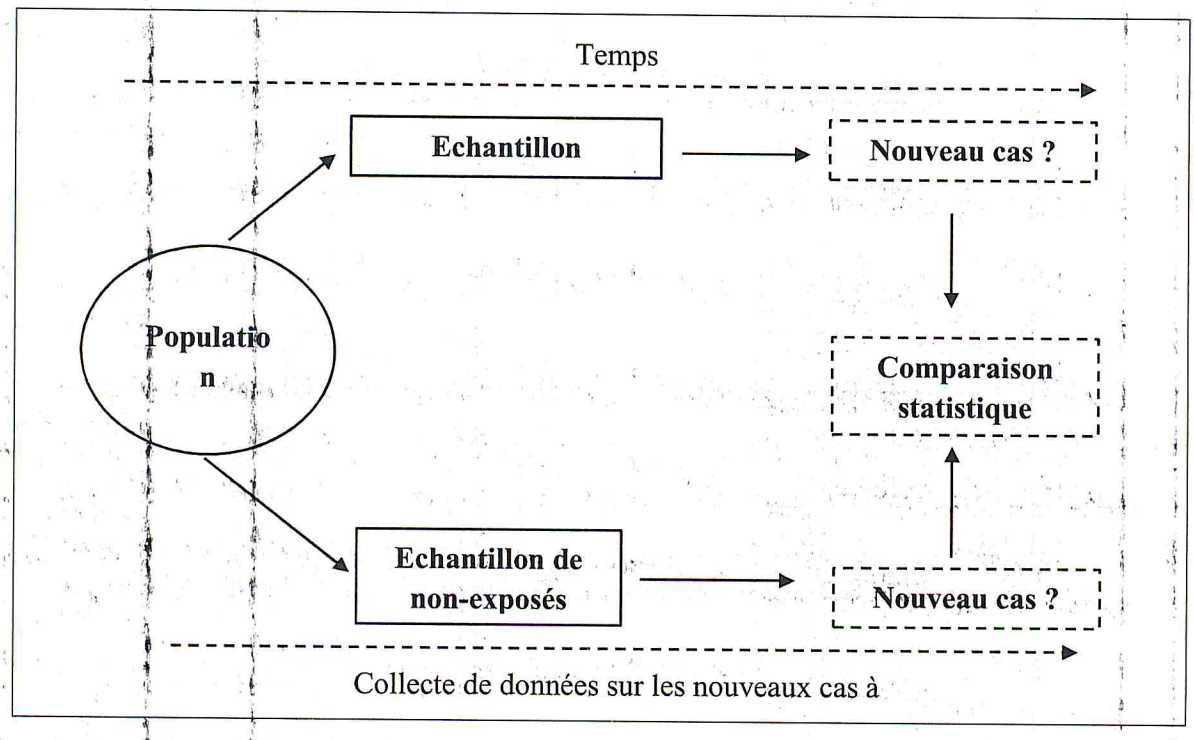


Figure 3. 2 – Schéma d'une enquête exposé non exposé [20]

3.2.2.2 Les enquêtes cas-témoins : [20]

Le principe de l'enquête cas-témoin est de comparer la fréquence de l'exposition antérieure à un facteur de risque dans un groupe de sujets malades (les cas) et dans un groupe de sujets témoins, indemnes de la maladie étudiée.

La planification d'une enquête cas-témoins commence par le choix de la population (dont seront issus les cas et les témoins) qui fera l'objet de l'enquête. Par définition, les cas seront atteints de la pathologie étudiée et représentatifs, pour l'exposition au facteur de risque, de l'ensemble des malades ayant cette pathologie. Le groupe témoin est construit pour servir de référence (ils sont représentatifs pour l'exposition au facteur de risque de la population dont sont issus les cas) et fournir une fréquence de base de l'exposition au facteur de risque dans la population dont sont issus les cas. Pour chacun des sujets inclus dans l'enquête (cas et témoins), des informations concernant l'exposition aux facteurs de risque vont être recherchées dans leur passé. On appelle souvent pour cette raison ces enquêtes des enquêtes rétrospectives. Différents modes de recueil sont utilisables : recherche dans les archives, interview des sujets, auto questionnaires.

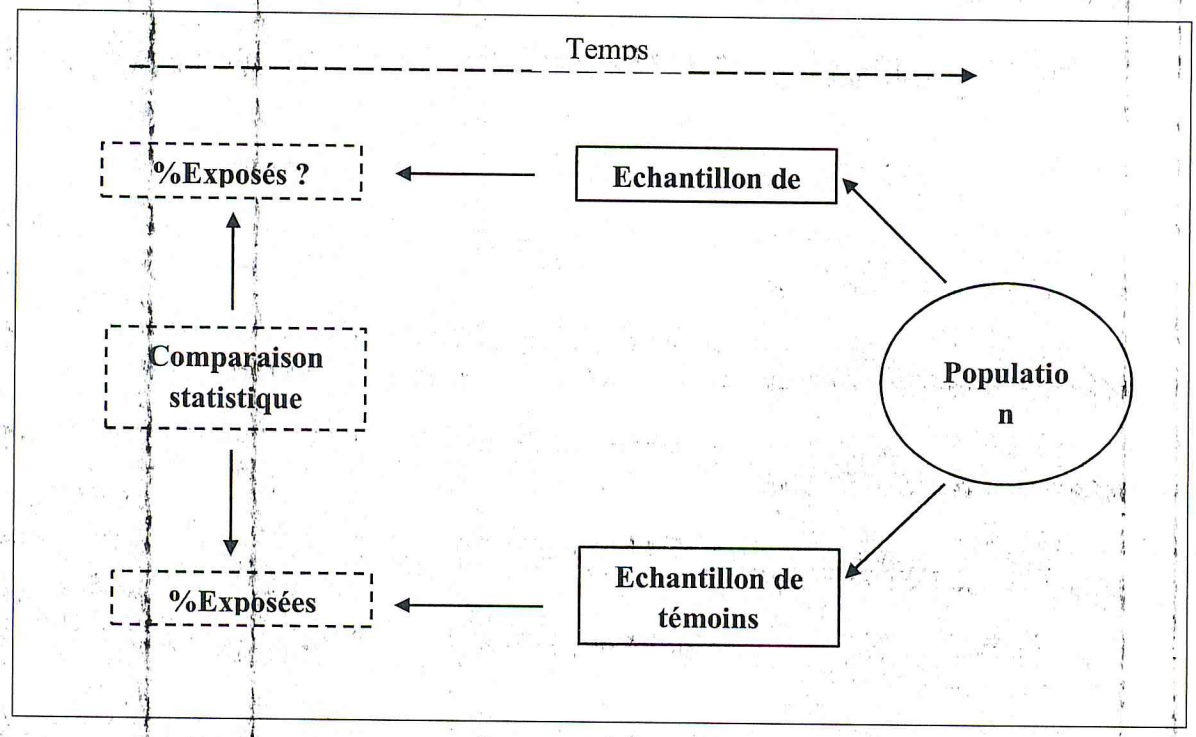


Figure 3. 3 – Schéma d'une enquête cas-témoins [20]

4. Notion de biais : [21]

Il s'agit d'erreurs systématiques survenues au cours de l'étude liées à l'investigateur, aux participants, ou au questionnaire par exemple, qui va induire une estimation d'un paramètre (moyenne, risque) qui va différer systématiquement de la valeur réelle, par excès ou par défaut. Il existe trois types de biais :

4.1 Biais de sélection : [22]

Les biais de sélection affectent la constitution de l'échantillon d'enquête, c'est à dire le processus par lequel les sujets sont choisis au sein de la population.

4.2 Biais de classement : [22]

Le biais de classement désigne une erreur systématique de mesure de l'exposition ou de la maladie. Ils conduisent à mal classer les sujets en « malades / non malades ».

4.3 Biais de confusion : [22]

Un biais de confusion désigne une erreur systématique dans l'estimation d'une mesure d'association (odds ratio ou risque relatif) entre le facteur étudié et la maladie, du fait d'un défaut de prise en compte d'un facteur de confusion.

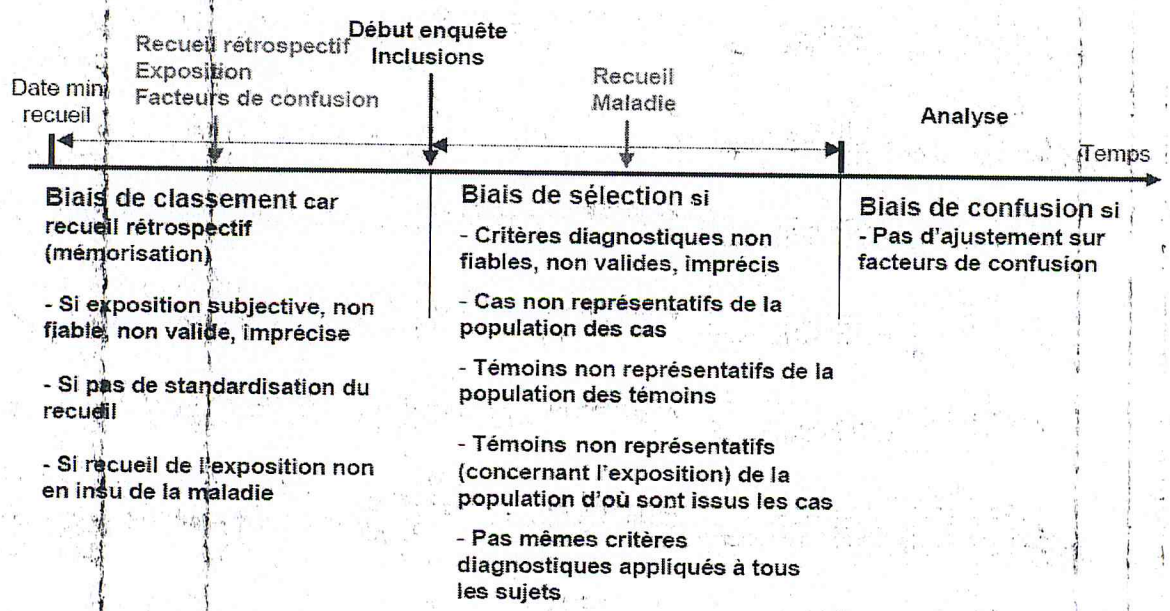


Figure 3. 4 - Sources de biais dans les enquêtes cas-témoins [20]

Partie 2 : Application de l'algorithme logistique pour la prédiction du cancer de sein

1. Introduction :

Le cancer est une préoccupation majeure de santé publique. Notre pays fait face à une croissance de l'incidence du cancer. Près de 40 000 cas de cette pathologie sont enregistrés annuellement sur le territoire national. Notamment les cancers du sein et des poumons, selon des chiffres de l'Institut national de santé publique (Insp). Les cancers du sein, du colon, des poumons, du col de l'utérus et de la prostate, demeurent les plus répandus en Algérie, avec un taux de prévalence de 50%, pour une moyenne d'âge de 59 ans chez l'homme et 51 ans chez la femme. Le taux important de prévalence de cette pathologie, durant les dernières années, est dû au changement du mode de vie, le vieillissement de la population et l'inadéquation du système sanitaire avec la démographie et la transition épidémiologiques importantes. Ainsi, une forte prévalence de la maladie a été enregistrée durant les dernières années, passant de 80 cas pour chaque 100.000 habitants en 1993 à plus de 120 cas pour chaque 100.000 habitant durant les dernières années avec un taux d'atteinte plus important chez les femmes. Les spécialistes ajoutent que seul le dépistage précoce permet de prévenir tous ces types de cancer et de réduire les cas de décès qui en résultent. Ils ont par ailleurs indiqué, et concernant certaines lacunes du système de santé, que des études ont démontré que 35% des protocoles en vigueur ne sont pas respectés par les médecins. Ils ont enfin plaidé pour le dépistage précoce en vue de réduire les risques d'atteinte de cette pathologie, donner aux malades la chance de traitement, accompagner ceux qui ont atteint un stade avancé de la maladie et réduire les frais de traitement [23].

Le cancer du sein est une tumeur maligne qui prend naissance dans les cellules du sein. Le mot « maligne » signifie que la tumeur peut se propager (métastases) vers d'autres parties du corps. Les cellules du sein subissent parfois des changements qui rendent leur mode de croissance ou leur comportement anormal. Ces changements peuvent engendrer des affections bénignes du sein, comme l'hyperplasie atypique et des kystes. Ils peuvent aussi entraîner la formation de tumeurs bénignes, dont les papillomes intracanalaires. Les affections et les tumeurs bénignes ne sont pas cancéreuses. Cependant, dans certains cas, des modifications dans les cellules mammaires peuvent causer un cancer du sein [24].

Le cancer du sein se développe le plus souvent dans les cellules qui tapissent les canaux, ou tubes, qui transportent le lait des glandes au mamelon. Ce type de cancer du sein est appelé carcinome canalaire. Le cancer peut aussi se former dans les cellules des glandes productrices de lait (regroupées en lobules). Ce type de cancer porte le nom de carcinome lobulaire. Ces deux cancers (carcinomes canalaire et lobulaire) peuvent être in situ, c'est-à-dire qu'ils demeurent confinés à leur emplacement d'origine et n'envahissent pas les tissus voisins. Ils peuvent également être infiltrants, ou invasifs, c'est-à-dire qu'ils se sont propagés dans les tissus voisins [24].

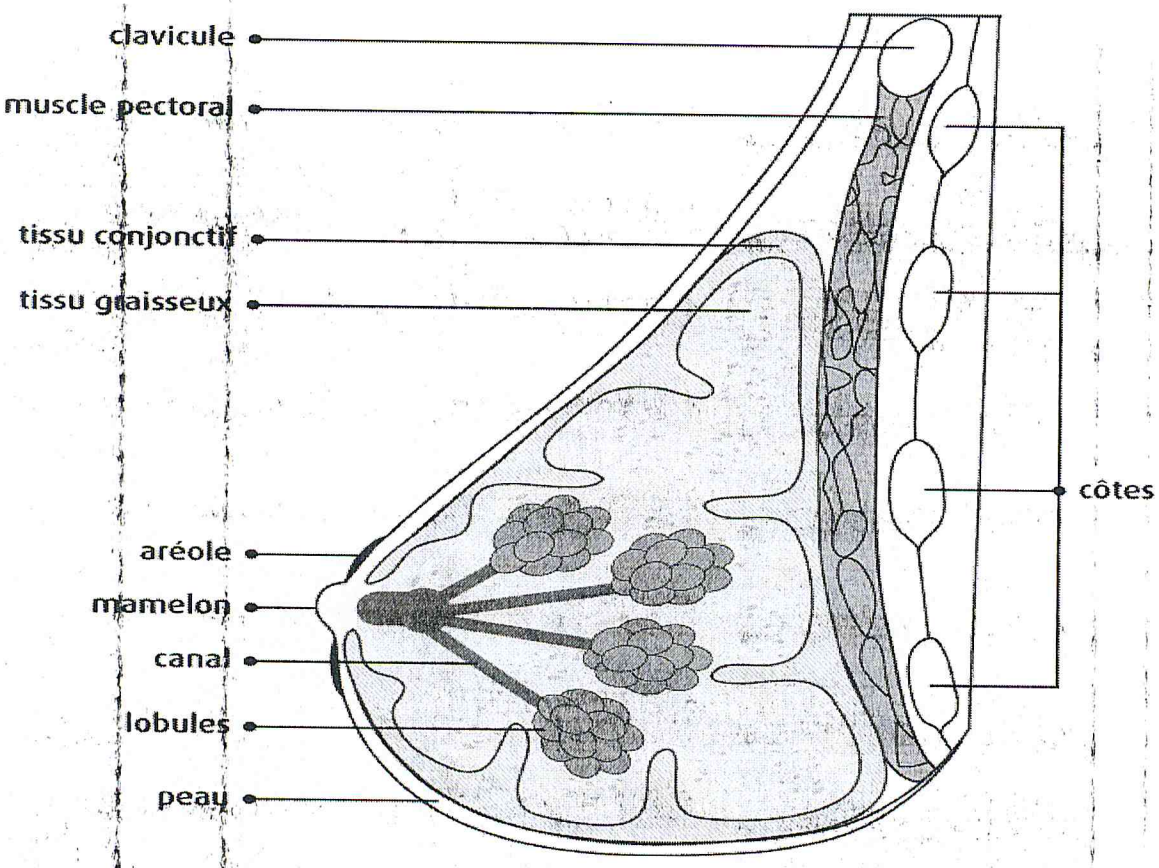


Figure 3. 5 – Anatomie du sein [24]

Cas d'étude 01 :

1. La population étudiée :

La population dont le comportement est à prédire dans le cadre de ce travail est celle des sujets atteints/non atteints d'un cancer de sein. Les données du travail présent sont issues d'une liste de fiches malades. Elles proviennent initialement du Centre Anti Cancer (CAC) de Blida et ont été collectées par le médecin spécialiste en cancérologie, Haffar Amina.

L'échantillon utilisé dans l'analyse est composé de 105 sujets dont 93 malades et 12 non malades. Il est à noter que le nombre des non malades est très faible en comparaison avec celui des malades, cette disparité est due au fait que la notion du dépistage n'est pas très répandue dans notre pays et les sujets reçus par le médecin spécialiste sont déjà malades pour la grande majorité. Il était donc pas évident de trouver un plus grand nombre de cas non malades.

2. Analyse de l'échantillon étudié (échantillon d'apprentissage):

Dans cette étude, la variable dépendante prend la valeur 1 quand il s'agit d'un sujet malade et 0 sinon. Les variables dépendantes sont les facteurs de risque liés à la maladie.

Après discussion avec le médecin du service, les facteurs de risque du cancer de sein sont les suivants :

- **Sexe :** le sexe du sujet.
- **Age :** l'âge du sujet.
- **Ménarchie précoce :** si les menstruations du sujets ont commencé à un jeune âge.
- **Mariage tardif :** si le sujet ne s'est pas marié à un jeune âge.
- **Allaitement :** s'il s'agit d'un allaitement excessif.
- **Contraception :** si le sujet utilise des contraceptifs.
- **Antécédent du cancer du sein :** si le sujet a des antécédents familiaux du cancer de sein.
- **Antécédent des autres cancers :** si le sujet a des antécédents familiaux d'autres cancers.
- **Habitudes alimentaires :** c'est l'influence de certaines habitudes alimentaires.
- **Résultats de la mammographie :** c'est les résultats de la mammographie.

Remarque : parmi les 105 sujets qu'on a pu mettre les mains dessus, on a trouvé 3 hommes atteints de cette maladie, ce qui prouve que la maladie n'est pas à 100% inexistante chez les hommes.

3. Description et codification des variables indépendantes :

Les facteurs de risque cités ci-dessus joueront par la suite le rôle des variables indépendantes et sont :

Sexe (X_1), Age (X_2), Ménarchie (X_3), Mariage (X_4), Allaitement (X_5), Contraception (X_6), Antécédent du cancer du sein (X_7), Antécédent des autres cancers (X_8), Habitudes alimentaires (X_9), Résultats de la mammographie (X_{10}). Le tableau ci-dessous (**Table 3.1**) présente les facteurs de risque de notre étude ainsi que leurs caractéristiques :

Variable	Définition	Caractéristique	Code	N(%)
X_1	Sexe	Femme Homme	1 0	96,2% 3,8%
X_2	Age	≥ 40 < 40	1 0	74,3% 25,7%
X_3	Ménarchie précoce	≤ 11 > 11	1 0	92,4% 7,6%
X_4	Mariage tardif	≥ 35 < 35	1 0	85,7% 14,3%
X_5	Allaitement	(+) (-)	1 0	80% 20%
X_6	Contraception	(+) (-)	1 0	75,2% 24,8%
X_7	Antécédent du cancer du sein	(+) (-)	1 0	98% 2%
X_8	Résultat de la mammographie	(+) (-)	1 0	77% 23%
X_9	Habitudes alimentaires	(+) (-)	1 0	75,8% 24,2%
X_{10}	Antécédent des autres cancers	(+) (-)	1 0	59% 41%

Table 3. 1 – Tableau de description des variables explicatives

Variable	Définition	Modalités	Effectifs	N(%)
Y	Maladie	0	12	11,429
		1	93	88,571

Table 3. 2 – Tableau de description de la variable expliquée

4. Analyse univariée - sélection des variables du modèle :

Comme étape première, on doit fixer nos variables, le choix de ces variables explicatives (X_i) n'est pas le fruit du hasard ni le résultat d'un screening fait à l'aveugle de plusieurs centaines de variables. Il est basé sur la connaissance de la physiopathologie de la maladie et les possibles facteurs influençant cette dernière. Ainsi, toutes les variables explicatives étudiées ne seront pas toutes nécessairement incluses dans l'analyse multivariée. En effet, ne seront introduites dans le modèle final que les variables qui pourraient avoir un lien avec la maladie.

4.1 Le rapport de cotes (OR) :

L'odds ratio quantifie l'association entre l'exposition à un facteur de risque et la maladie étudiée. Il varie entre zéro et l'infini, quantifie la force de l'association entre la survenue de l'événement représenté par la variable dichotomique et les variables explicatives [25].

	Cas	Témoin
Exposition au facteur (E+)	a	b
Non exposition au facteur (E-)	c	d

Table 3. 3 – Tableau de contingence

Il se calcule comme suit :

OR = (a/c) / (b/d) = (a*d) / (b*c).

En l'absence d'association entre l'exposition à un facteur de risque et la maladie en question, il tend vers 1, à l'inverse lorsque les variables sont fortement liées, il tend vers zéro ou l'infini [25].

Ainsi on note :

OR = 1 : facteur neutre (absence d'association)

0 < OR < 1 : facteur protecteur

OR > 1 : facteur de risque

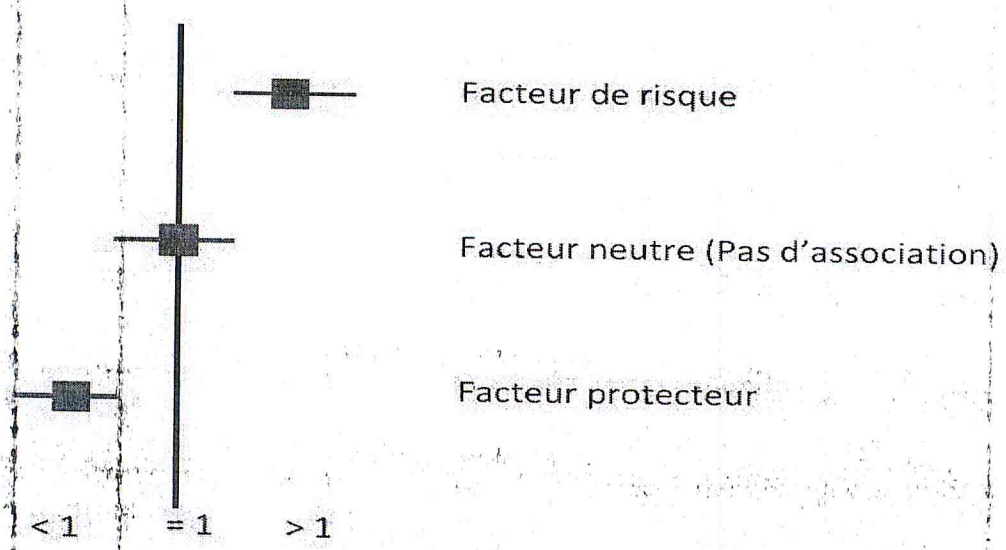


Figure 3. 6 – Diagramme d'interprétation de l'OR [26]

4.2 Intervalle de confiance (IC95%) :

C'est l'intervalle de valeurs entre lesquelles les valeurs réelles de l'OR ont 95% de chances de se retrouver dans la population étudiée. Par définition, si l'IC contient la valeur 1 alors l'augmentation du risque évaluée par l'OR [26] est « non significative », du coup :

$1 \notin$ IC95% →Association statistiquement significative entre l'exposition et la maladie.

$1 \in$ IC95% →L'exposition n'est pas statistiquement liée à la maladie.

Elle est calculée comme suit :

BI (borne inférieure) = $\ln(OR) - 1{,}96(1/A+1/B+1/C+1/D)^{1/2}$

BS (borne supérieure) = $\ln(OR) + 1{,}96(1/A+1/B+1/C+1/D)^{1/2}$

IC95% de l'OR = [exp(BI) ; exp(BS)] = $e^{\ln(OR) \pm 1{,}96(1/A+1/B+1/C+1/D)1/2}$

4.3 Test de CHI² (Khi-deux) : [26]

Le test de CHI² (χ^2) est un test permettant de vérifier s'il existe une relation entre le risque d'exposition et la maladie. Il se calcule de la manière suivante :

$$\chi^2 = \sum R^2 / T_0$$

Où $\mathbf{R = T - T_0}$ (T = effectifs observés et T_0 = effectifs théoriques)

Si on prend comme exemple le calcul de χ^2 pour la variable X_5 qui correspond à l'allaitement, on procède ainsi :

Étape 1: Poser les hypothèses.

Ho : Les variables sont indépendantes ⇒ OR= 1.

H1 : Les variables ne sont pas indépendantes ⇒ OR ≠ 1.

	Cas	**Témoin**	**Total**
Oui	78	6	84
Non	15	6	21
Total	93	12	105

Table 3. 4 – Tableau d'effectifs observés T

Étape 2 : Calculer le tableau des effectifs théoriques ou tableau d'indépendance que l'on appelle T_0 grâce à la formule suivante :

$\mathbf{T_0}$ = (total de la ligne) * (total de la colonne) / total de l'échantillon.

$$\frac{21*93}{105} = 18.6$$

	Cas	**Témoin**	**Total**
Oui	74.4	9.6	84
Non	18.6	2.4	21
Total	93	12	105

Table 3. 5 – Tableau d'effectif théorique T_0

Enfin, on divise termes à termes le tableau $R^2 (T - T_0)$ par le tableau des effectifs théoriques T_0 et on aura :

$\chi^2_{\text{allaitement}} = \sum R^2 / T = 7.62.$

4.4 Le degré de significativité P : [27]

La « valeur P » est une mesure statistique qui aide les scientifiques à déterminer si leurs hypothèses sont correctes. Celle-ci est utilisée pour savoir si les résultats d'une expérience se trouvent dans la gamme normale des valeurs pour un événement observé. Généralement, si la valeur P d'un jeu de données est au-dessous d'une valeur prédéterminée (comme 0,20 par exemple), les scientifiques rejetteront l' « hypothèse nulle » de leur expérience – autrement dit, ils élimineront l'hypothèse selon laquelle les variables testées au cours de leur expérience n'ont *aucun* effet significatif sur les résultats. Les valeurs P sont lues dans un tableau de référence après avoir calculé la valeur du Khi-deux.

La valeur de P est obtenue selon la valeur de χ^2 et le degré de liberté DDL, elle se calcule comme suit :

DDL = (nombre de lignes – 1) * (nombre de colonnes).

De ce fait, le DDL dans notre étude est égale à 1.

p	0,999	0,995	0,99	0,98	0,95	0,9	0,8	0,2	0,1	0,05	0,02	0,01	0,005	0,001
ddl														
1	0,0000	0,0000	0,0002	0,0006	0,0039	0,0158	0,0642	1,6424	2,7055	3,8415	5,4119	6,6349	7,8794	10,8276
2	0,0020	0,0100	0,0201	0,0404	0,1026	0,2107	0,4463	3,2189	4,6052	5,9915	7,8240	9,2103	10,5966	13,8155
3	0,0243	0,0717	0,1148	0,1848	0,3518	0,5844	1,0052	4,6416	6,2514	7,8147	9,8374	11,3449	12,8382	16,2662

Table 3. 6 – Tableau de Khi-Deux

Si la valeur P est plus faible que le seuil de significativité fixé dans notre étude à 0.20, alors on rejette hypothèse nulle (H_0) de l'indépendance et nous concluons que la variable explicative étudiée (X_i) et la variable à expliquer (Y) sont dépendants.

Les différents calculs de cette analyse sont présentés dans le tableau ci-dessous :

Analyse univariée		Cas	Témoin	OR/Brut	IC95%	X^2	P
Sexe	*Féminin*	90	11	2.7	]0.26,28.54[	0.76	**0.40**
	masculin	3	1				
Age	*>= 40*	70	8	1.52	]0.42,5.52[	0.41	**0.52**
	< 40	23	4				
Ménarchie	*Oui*	90	7	21.43	]4.22,108.82 [	22.31	**0.0002**
	Non	3	5				
Mariage	*Oui*	83	7	5.93	]1.58,22.24[	8.3	**0.0083**
	Non	10	5				
Allaitement	*Oui*	78	6	5.2	]1.48,18.32 [	7.62	**0.01**
	Non	15	6				
Contraception	*Oui*	74	5	5.45	]1.65,19.09 [	8.2	**0.008**
	Non	19	7				
Antécédent du cancer du sein	*Oui*	60	3	5.45	]1.38,21.53 [	6.92	**0.015**
	Non	33	9				
Antécédent des autres cancers	*Oui*	21	3	0.88	]0.22,3.55 [	0.04	**0.85**
	Non	72	9				
Habitudes alimentaires	*Oui*	0	0	0.13	]0.002,7.04 [	-	**0.32**
	Non	93	12				
Résultats de la mammographie	*Oui*	91	10	9.1	[1.15, 71.82]	6.11	**0.03**
	Non	2	2				

Table 3. 7 – Tableau des résultats de l'analyse univariée

4.5 Les méthodes pas à pas :

En statistiques, il existe ce qu'on appelle des analyses *« pas à pas »*. Deux stratégies, à peu près équivalentes sont possibles : ascendante ou descendante.

Dans la méthode ascendante [25] : on introduit en premier la covariable la plus significativement associée à la variable expliquée. Puis on étudie à chaque pas l'impact sur l'explication du pronostic de l'introduction dans le modèle de la covariable suivante. Bien entendu, on ne garde en définitive que les covariables associées à une significativité $\leq 0,05$ au sein du modèle multivarié.

Dans cette méthode, on procède à l'utilisation de l'algorithme suivant :

- *On estime d'abord les paramètres pour les variables présentes dans le modèle.*
- *On calcule ensuite pour chaque variable non présente dans le modèle, la statistique du « Khi-deux », c'est-à-dire la statistique du score pour le test :*
  ***Ho :** modèle comprenant toutes les variables entrées jusqu'à cette étape.*
  ***H1 :** modèle comprenant toutes les variables entrées jusqu'à cette étape plus la variable examinée.*
- *Si une de ces statistiques est significative (par défaut, 0.05), la variable pour laquelle la statistique est la plus grande est entrée dans le modèle. On revient à l'étape d'estimation pour le modèle augmenté.*
- *Sinon, la procédure est terminée, et le modèle retenu est celui de la dernière étape.*

Dans la méthode descendante [25] : toutes les covariables sont introduites d'emblée dans le modèle. A chaque pas, on élimine une variable du modèle, de façon à ce qu'il ne reste lors du dernier pas que les covariables associées significativement *($p \leq 0,05$)* à la variable expliquée *(le pronostic)*.

L'algorithme suivant résume les étapes de cette méthode :

(1) Dans cette méthode, on estime les paramètres pour les variables encore présentes dans le modèle.
On passe en (2).
(2) Si toutes les variables sont significatives individuellement (par défaut, 0.05), la procédure s'arrête.
- Si une des variables n'est pas significative individuellement, la moins significative est éliminée du modèle. On passe en (1).

Le diagramme suivant résume la démarche suivi dans cette étape :

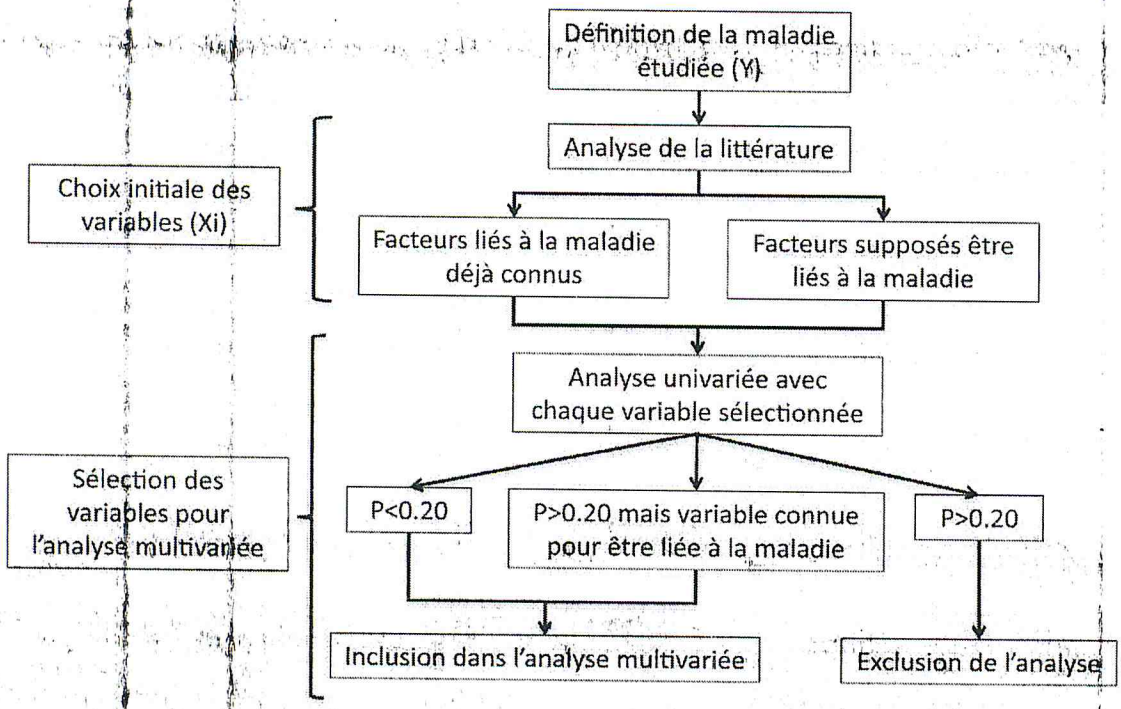


Figure 3. 7 – Diagramme expliquant le processus de sélection des variables explicatives à inclure dans un modèle de régression logistique [7]

Les variables explicatives qui sont liées de façon suffisamment forte à la variable à expliquer sont alors conservées dans le modèle. D'une manière générale, toutes les variables dont le degré de significativité est inférieur à 0,20 (c'est-à-dire avec une valeur de $p < 0,20$) lors de l'analyse univariée seront incluses dans le modèle initial de régression logistique multiple. Bien entendu un tel seuil peut sembler arbitraire et peut varier selon les habitudes des différentes équipes (il est parfois de 0,25, parfois de 0,30). Il est important de signaler que seront aussi incluses dans l'analyse des variables connues pour être associées à la maladie, même si l'analyse univariée n'a pas abouti à une valeur de $p < 0,20$. On parle de variables dites « forcées ». Il apparaît donc clairement que le choix des variables à inclure est avant tout basé sur une réflexion clinique (variables connues comme étant associées à la maladie) et prend en compte des arguments statistiques.

4.6 Résultats :

On a vérifié au préalable par l'analyse précédente, l'impact des facteurs de risque : sexe, âge, ménarchie précoce, mariage tardif, allaitement, contraception, antécédents du cancer de sein, antécédents d'autres cancers, habitudes alimentaires, mammographie sur la survenue du cancer de sein. Seuls la ménarchie, mariage, allaitement, contraception, antécédents du cancer de sein, résultat de mammographie présentent une influence sur l'apparition de la maladie en

question. Par contre, les variables X_8 et X_9 correspondantes respectivement aux antécédents des autres cancers et aux habitudes alimentaires, elles ne seront pas prises en compte dans notre modèle par le fait qu'elles ont présenté des résultats non significatifs comme nous le démontre le tableau ci-dessus (**Tableau 3.7**).

Quant aux variables explicatives : sexe et âge, la valeur de P étant nettement supérieure à notre seuil fixé, donc statistiquement parlant, elles ne seront pas incluses dans notre modèle. Cependant, d'un point de vue clinique, elles sont jugées indispensables et doivent y figurer, du coup, elles seront « forcées » dans notre modèle.

5. Analyse multivariée :

La régression logistique permet d'étudier la relation entre une variable réponse binaire et plusieurs variables explicatives. Comme dans la régression linéaire, on cherche la meilleure combinaison linéaire des données d'entrée pour modéliser la réponse, à ceci près que c'est une transformation de cette combinaison (on parle alors d'*une fonction de lien : LOGIT*) qui est utilisée en sortie.

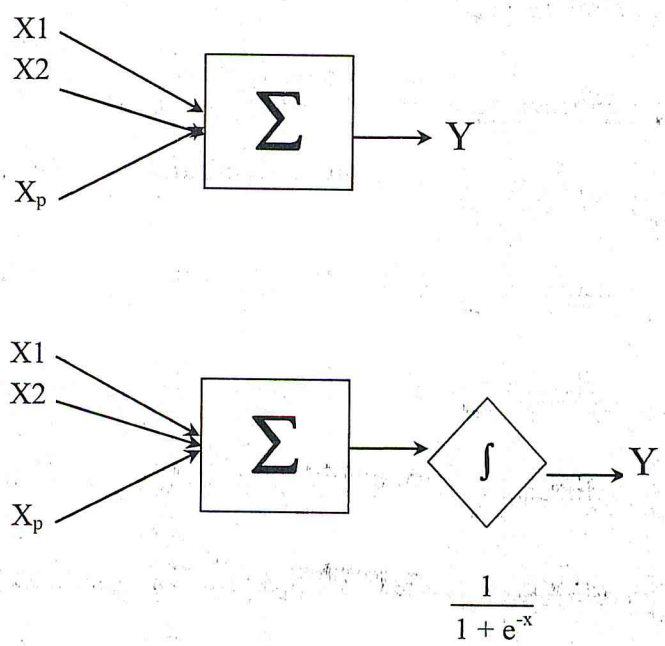


Figure 3. 8 – Schéma de représentation de la fonction de lien

L'objectif est de prédire et/ou expliquer une variable catégorielle Y à partir d'une collection de descripteurs $X = (X_1, X_2, \ldots, X_j)$. Il s'agit en quelque sorte de mettre en évidence l'existence d'une liaison fonctionnelle sous-jacente (en Anglais, underlying concept) de la forme : $Y = f(X,\alpha_i)$ entre ces variables. La fonction f est le modèle de prédiction, on

parle aussi de classifieur, α est le vecteur de paramètres de la fonction, on doit en estimer les valeurs à partir des données disponibles [31].

Dans notre cas, nous considérons que la variable dépendante Y ne prend que 2 modalités (variable binaire) : la valeur « 1 » ou la valeur « 0 ». Chacune de ces valeurs correspond à une classe prédéfinie (ici la classe « malade » ou la classe « non malade). Nous cherchons alors à prédire correctement les valeurs de Y.

Lorsque le modèle final est atteint (plus petit nombre possible de variables explicatives liées significativement à la variable dépendante), on vérifie l'absence d'interaction entre les variables du modèle final. L'analyse de corrélation permet de quantifier la force du lien entre des variables.

Le but des analyses multivariées est de sélectionner, parmi l'ensemble des liaisons statistiques mises en évidence par les analyses univariées, la ou les covariables qui expliquent de façon indépendante la survie ou la maladie.

L'analyse multivariée peut mettre en évidence des liens entre covariables et éliminer progressivement celles qui ne sont pas des facteurs prédictifs indépendants entre eux (de la maladie ou de l'évolution) et qu'on désigne sous le terme de facteurs de confusion. Elle décante les variables pour ne retenir au final que celles qui suffisent à expliquer sans redondance le modèle.

Variables	age	menarchie	mariage	allaitemen	ntraceptio	anteceden	sexe	amograph
age	**1,000**	-0,005	0,445	0,468	0,218	-0,047	-0,117	-0,082
menarchi	-0,005	**1,000**	0,088	0,395	0,417	0,199	0,693	-0,040
mariage	0,445	0,088	**1,000**	0,680	0,333	-0,008	-0,081	0,142
allaitemer	0,468	0,395	0,680	**1,000**	0,541	0,068	0,398	0,105
contracep	0,218	0,417	0,333	0,541	**1,000**	0,150	0,347	0,243
anteceder	-0,047	0,199	-0,008	0,068	0,150	**1,000**	0,138	0,026
sexe	-0,117	0,693	-0,081	0,398	0,347	0,138	**1,000**	-0,028
mamogra	-0,082	-0,040	0,142	0,105	0,243	0,026	-0,028	**1,000**

Table 3. 8 - Matrice de proximité (Coefficient de corrélation de Pearson)

6. Résultats :

On observe que la variable Age possède un lien inverse et très faible avec les variables ménarchie, Antécédent du cancer de sein, Sexe, et Mammographie, ce qui signifie que si une variable augmente, l'autre diminue. Aussi, l'Âge a un lien directe faible avec la variable

Contraception et qui veut dire si une variable augmente, l'autre augmente aussi. De même, il existe un lien modéré avec les variables Mariage et Allaitement.

La variable Ménarchie a un lien inverse et très faible avec les variables Âge et Mammographie, tandis qu'elle a un lien directe et très faible avec les variables Mariage et Antécédent du cancer de sein et un lien directe faible avec la variable Allaitement aussi. Elle a un lien direct modéré avec la variable Contraception, mais un lien fort avec la variable Sexe.

La variable mariage possède un lien inverse faible avec les variables Antécédent du cancer de sein et Sexe. Par contre elle a un lien direct très faible avec les variables Ménarchie et Mammographie. Aussi, elle possède un lien faible avec la variable Contraception et un lien direct modéré avec la variable Âge et un lien direct fort avec la variable Allaitement.

La variable allaitement a un lien directe très faible avec les variables Antécédent du cancer de sein et Mammographie ainsi d'un lien faible avec les variables Ménarchie et Sexe, et un lien modéré avec les variables Âge et Contraception. De plus, elle possède un lien fort avec la variable Mariage.

La variable Contraception a un lien direct très faible avec la variable Antécédent du cancer de sein et un lien faible avec les variables Sexe, Mammographie, Mariage et Âge, tandis qu'elle possède un lien modéré avec les variables Allaitement et Ménarchie.

La variable Antécédent du cancer de sein a un lien inversé très faible avec les variables Âge et Mariage et un lien direct très faible avec les variables Ménarchie, Allaitement, Sexe, Mammographie et Contraception.

La variable Sexe possède un lien inversé très faible avec les variables Âge, Mariage et Mammographie. Tandis qu'elle a un lien direct très faible avec les variables Ménarchie, Allaitement, Contraception et Antécédent du cancer de sein.

La variable Mammographie a un lien très faible avec toutes les variables, sauf qu'avec les variables Âge, Ménarchie et Sexe le lien est inversé.

7. Le modèle estimé :

Les paramètres de notre modèle ont été estimés par la méthode du maximum de vraisemblance à l'aide du logiciel STATISTICA. Le tableau suivant les résume :

Constante	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
-0.3026	-0.0034	0.8706	0.1016	0.0507	-0.0552	0.1435	-0.7125	0.9186

Table 3. 9 - Tableau des valeurs des coefficients estimées par le logiciel STATISTICA

8. Validation du modèle :

Après avoir estimé les paramètres de notre modèle, on procède à la validation de ce dernier.

8.1 Critères de validation d'un modèle de régression logistique : [36]

- **Performance :**

Discrimination : capacité à différencier les "malades" des "non malades" → surface ROC.

- **Calibration :**

Concordance entre les probabilités prédites et observées Exp : test de Hosmer et Lemeshow.

- **Mesures globales :**

Exp : R^2 de Nagelkerke, R squared, R Mcfadden, R de Cox & Snell

Afin d'évaluer la qualité du modèle estimé, nous utilisons plusieurs statistiques et tests. Nous vérifions également dans quelle mesure les prédictions du modèle correspondent à la réalité observée. Pour cela, le logiciel XLSTAT propose des statistiques comme le R-deux de Cox & Snell (basé sur la log-vraisemblance du modèle comparée avec celle d'un modèle constant ; il prend toujours des valeurs inférieures à 1, même pour un modèle « parfait ») ou de Nagelkerke (c'est une version ajustée du R-deux de Cox & Snell, qui prend des valeurs dans l'intervalle [0 ;1]), mais aussi d'autres tests, comme celui de Hosmer-Lemeshow (plus robuste, en raison du fait qu'il est basé sur le regroupement des observations en déciles et la comparaison de la probabilité observée avec la probabilité théorique à l'intérieur de chaque décile).

8.2 Le coefficient de détermination R^2 : [28]

Le coefficient de détermination (R^2) est un indicateur qui permet de juger la qualité d'une régression linéaire, simple ou multiple. D'une valeur comprise entre 0 et 1, il mesure l'adéquation entre le modèle et les données observées. Il permet de connaître l'intensité de la relation unissant deux paramètres X et Y.

$$R^2 = 1 - \frac{\ln L_\beta}{\ln L_\alpha}$$

L_0 et L_B représentent respectivement la vraisemblance du modèle initial logit P0(X) = α et du modèle d'intérêt logit P(X) = Xβ et *n* correspond à la taille de l'échantillon.

Exemple : R^2 = 35% signifie que 35% des variations de la variable dépendante sont expliqués par le modèle de régression et que 65% restent par conséquent inexpliqués. La racine carrée du coefficient R^2 donne le coefficient de corrélation. Le R^2 ajusté est utilisé en cas de régression multiple. Il s'interprète de la même manière que le R^2 mais tient compte de l'augmentation du nombre de variables explicatives.

$$R^2_{\text{ajusté}} = 1 - \frac{\ln L_\beta - K}{\ln L_\alpha}$$

Où K = nombre de variables explicatives.

Le tableau suivant indique la valeur de R^2 de notre étude :

R^2(McFad	0,000	0,616		
R^2(Cox an	0,000	0,355	R^2	0,547
R^2(Nagelk	0,000	0,697	R^2 ajusté	0,510

Table 3. 10 - Tableau des valeurs du coefficient de détermination estimées par le logiciel STATISTICA

8.3 Test de Hosmer et Lemeshow : [35]

Le principe du test de Hosmer et Lemeshow consiste à comparer les valeurs prédites et observées des modalités de la variable d'intérêt, après regroupement des individus en classes. On utilise ensuite la distance de Khi-deux pour calculer la distance entre les fréquences observées et prédites. Lorsque cette distance est relativement petite, on considère que le modèle est bien calibré. Le test repose sur les hypothèses suivantes :

H_0 : le modèle est bien calibré contre H1 : le modèle n'est pas bien calibré.

La lecture du tableau suivant relatif aux résultats du test d'Hosmer et Lemeshow montre que l'ajustement global du modèle aux données est satisfaisant. Car, la valeur de la probabilité critique (Prob > chi2) est supérieure au seuil de signification de 5%.

Test de Hosmer-Lemeshow (Variable class			
Statistique	Khi²	DDL	Pr > Khi²
Statiqtiqu	20,023	8	0,010

Table 3. 11 - Tableau des résultats du test de Hosmer et Lemeshow comme retournés par le logiciel STATISTICA

8.4 La courbe ROC : [30]

La fonction d'efficacité du récepteur, plus fréquemment désignée sous le terme « courbe ROC » (de l'anglais Receiver Operating Characteristic), dite aussi caractéristique de performance (d'un test) ou courbe sensibilité/spécificité, est une mesure de la performance d'un classificateur binaire, c'est-à-dire d'un système qui a pour objectif de catégoriser des éléments en deux groupes distincts sur la base d'une ou plusieurs des caractéristiques de chacun de ces éléments. Graphiquement, on représente souvent la mesure ROC sous la forme d'une courbe qui donne le taux de vrais positifs en fonction du taux de faux positifs.

On étudie alors un ensemble de valeurs seuil possibles et, pour chacune, on calcule différentes statistiques dont les plus simples sont :

- Vrais positifs (**VP**) : nombre d'individus déclares positifs par le test et qui le sont effectivement.
- Faux positifs (**FP**) : nombre d'individus déclares positifs par le test mais qui sont en réalité négatifs.

- Vrais négatifs (**VN**) : nombre d'individus déclares négatifs par le test et qui le sont effectivement.
- Faux négatifs (**FN**) : nombre d'individus détectes négatifs par le test mais qui sont en réalité positifs.
- Prévalence de l'évènement : fréquence de survenance de l'évènement dans l'échantillon total (**VP+FN)/N**.
- *Sensibilité* (aussi appelée Fraction de Vrais Positifs): proportion d'individus positifs effectivement bien détectes par le test. Autrement dit, la sensibilité permet de mesurer à quel point le test est performant lorsqu'il est utilisé sur des individus positifs. Le test est parfait pour les individus positifs lorsque la sensibilité vaut 1, équivalent à un tirage au hasard lorsque la sensibilité vaut 0.5. S'il est inférieur à 0.5, le test est contre-performant.

La définition mathématique est : **Sensibilité = VP/(VP + FN)**.

- *Spécificité* (aussi appelée Fraction de Vrais Négatifs): proportion d'individus négatifs effectivement bien détectes par le test. Autrement dit, la spécificité permet de mesurer à quel point le test est performant lorsqu'il est utilisé sur des individus négatifs. Le test est parfait pour les individus négatifs lorsque la spécificité vaut 1, équivalent à un tirage au hasard lorsque la spécificité vaut 0.5. S'il est inférieur à 0.5, le test est contre-performant.

La définition mathématique est : **Spécificité = VN/(VN + FP)**.

L'aire sous la courbe (ou *Area Under the Curve – AUC*) est un indice synthétique calculé pour les courbes ROC. L'AUC correspond à la probabilité pour qu'un événement positif soit classé comme positif par le test sur l'étendue des valeurs seuil possibles. Pour un modèle idéal, on a AUC=1 (ci-dessus en bleu), pour un modèle aléatoire, on a AUC=0.5 (ci-dessus en rouge). On considère habituellement que le modèle est bon dès lors que la valeur de l'AUC est supérieure à 0.7.

Un modèle bien discriminant doit avoir une AUC entre 0.87 et 0.9. Un modèle ayant une AUC supérieure à 0.9 est excellent.

La figure à suivre représente les résultats de l'analyse de ROC qu'on a appliqué sur nos données d'étude en utilisant l'outil d'analyse statistique XLSTAT :

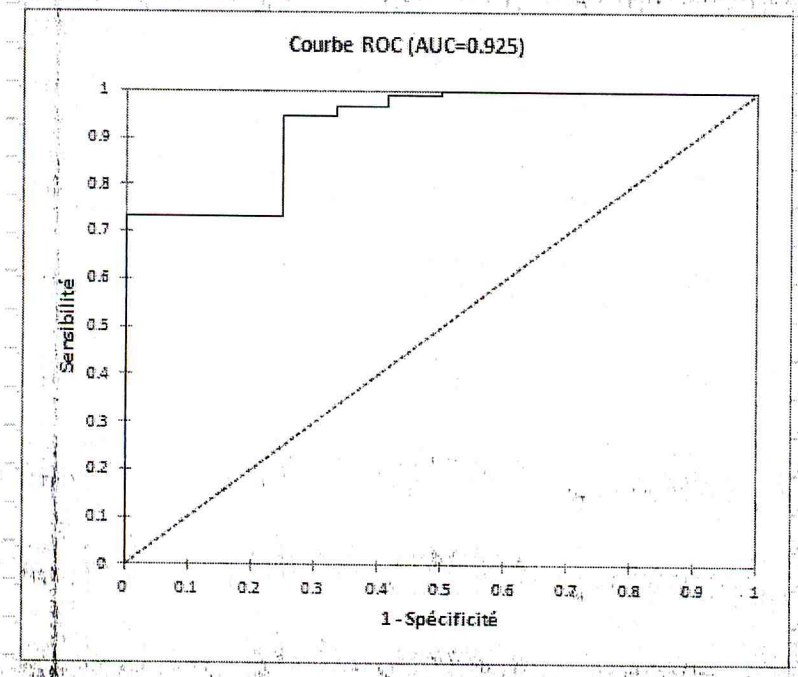


Figure 3. 9 – Tracé représentant les résultats de l'analyse de ROC

Selon la valeur d'AUC qu'est considéré comme un moyen de jugement de la qualité du modèle obtenu, notre modèle est qualifié d'excellent (AUC > 0.90).

Cas d'étude 02 :

1. La population étudiée :

Dans le cas présent, l'étude a été conduite dans le but de prédire les cellules ayant une croissance anormale ou proprement dit, une tumeur maligne ou bénigne. Nous avons utilisé la base de données du cancer du sein dénommée « Wisconsin Breast Cancer Database » qui a été obtenue par l'Université du Wisconsin et est publiée sur le site web : http://archive.ics.uci.edu/ml/datasets/.

2. Analyse de l'échantillon étudié (échantillon d'apprentissage):

Notre base de données contient les informations médicales de 699 cas cliniques relatifs au cancer du sein classés comme bénin ou malin : 458 patientes (soit 65.5%) sont des cas bénins et 241 patientes (soit 34.5%) sont des cas malins. Pour commencer, nous avons divisé l'ensemble des sujets en deux échantillons : un échantillon d'apprentissage S_A et un échantillon de test S_T. Le premier échantillon permet de générer le modèle logistique. Le deuxième échantillon c'est l'échantillon test qui a pour objectif de vérifier si le modèle fondé sur l'échantillon d'apprentissage est statistiquement fiable. La base de données contient 16 données manquantes, les facteurs de risque sont les suivants :

- **Clump Thickness :** l'épaisseur de la membrane plasmique d'une cellule cancéreuse est plus importante que celle d'une cellule normale.
- **Uniformity of Cell Size :** les cellules cancéreuses sont caractérisées par une anisocytose, à savoir une inégalité au niveau de la taille par comparaison avec les cellules saines.
- **Uniformity of Cell Shape :** les cellules cancéreuses sont marquées par des contours irréguliers ainsi que des incisures.
- **Shape Marginal Adhesion :** une surexpression de la protéine integrin beta3 au niveau de la surface de la cellule cancéreuse.
- **Single Epithelial Cell Size :** étant donné que les cellules épithéliales sont absentes à l'état naturel au niveau de la moelle osseuse et qu'elles ne sont pas détectées chez les individus sains, la moelle osseuse peut, de ce fait, être considérée comme un indicateur de maladie métastatique chez les patients atteints du cancer du sein au stade primaire.

- **Bare Nuclei :** à l'état normal, les nucléoles se trouvent à l'intérieur du noyau. Dans le cas où ces derniers se trouvent confondus avec le cytoplasme cela indique que la cellule présente une anomalie et qu'elle est susceptible de devenir cancéreuse.
- **Bland Chromatin :** H2az est une protéine qui induit l'expression du gène du récepteur d'oestrogènes. La surproduction de cette protéine est un marqueur de présence de cellules cancéreuses au niveau du sein étant donné qu'elles son hormono-dépendantes.
- **Normal Nucleoli :** L'ADN est naturellement protégé par une membrane nucléraire. Une défaillance observée au niveau de cette membrane peut refléter une croissance tumorale.
- **Mitoses :** La mitose est un processus de division cellulaire régulé permettant de reproduire des cellules filles génétiquement identique à la cellule parentale. Les cellules malignes sont caractérisées par une division cellulaire anarchique et intense par comparaison avec une population cellulaire normale.

Remarque : Etant donné qu'il y a 16 données manquantes, nous nous sommes restreints à travailler sur 683/699 patientes.

3. Description et codification des variables indépendantes :

La variable dépendante dans cette étude c'est la probabilité que la tumeur soit maligne et les variables indépendantes sont :

Clump_Thickness (X_1), Uniformity_of_Cell_Size (X_2), Uniformity_of_Cell_Shape (X_3), Marginal_Adhesion (X_4), Single_Epithelial_Cell_Size (X_5), Bare_Nuclei (X_6), Bland_Chromatin (X_7), Normal_Nucleoli (X_8), Mitoses (X_9).

4. Analyse univariée :

Le tableau ci-dessous résume les résultats de l'analyse univariée :

Analyse univariée		IC 95%	OR	X^2	P
Clump Thickness	1 - 10	]0.15, 0.24[	1.74	8,588	**<0,0001**
Uniformity of Cell Size	1 - 10	]0.016, 0.18[	0.85	2,324	**0,0205**
Uniformity of Cell Shape	1 - 10	]0.020, 0.18 [	1.40	2,447	**0,0147**
Marginal Adhesion	1 - 10	]-0.023, 0.080[	1.38	1,096	0,2734
Single Epithelial Cell Size	1 - 10	]0.006, 0.10[	1.06	2,183	**0,0294**
Bare Nuclei	1 - 10	]0.32, 0.42[	1.65	13,840	**<0,0001**
Bland Chromatin	1 - 10	]0.026, 0.13[	1.48	2,888	**0,0040**
Normal Nucleoli	1 - 10	]0.074, 0.17[	1.39	4,762	**<0,0001**
Mitoses	1 - 10	]-0.014, 0.062[	2.10	1,250	0,2119

Table 3. 12 – Tableau des résultats de l'analyse univariée sur l'échantillon d'apprentissage

4.1 Résultats :

L'analyse univariée de notre échantillon a permet de conclure que les variables que notre modèle prendra en compte sont : Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli car comme l'indique le tableau précédent, leur valeur de p est < 0,0001.

5. L'analyse multivariée :

Matrice de corrélation :

Variables	mp Thickn	mity of Cell	nity of Cel	pithelial	Bare Nucl	end Chrom	rmal Nucleo
Clump Thi	**1,000**	0,659	0,672	0,512	0,578	0,559	0,534
Uniformit	0,659	**1,000**	0,906	0,734	0,701	0,754	0,725
Uniformit	0,672	0,906	**1,000**	0,703	0,726	0,733	0,722
Single Epi	0,512	0,734	0,703	**1,000**	0,578	0,618	0,631
Bare Nucl	0,578	0,701	0,726	0,578	**1,000**	0,701	0,609
Bland Chr	0,559	0,754	0,733	0,618	0,701	**1,000**	0,686
Normal Nu	0,534	0,725	0,722	0,631	0,609	0,686	**1,000**

Table 3. 13 - Matrice de proximité (Coefficient de corrélation de Pearson)

On observe que la variable Clump_thickness a un lien directe avec les variables UCS et UCSH, ce qui signifie que si la valeur de CT augmente, les valeurs de UCS et UCSH augmente aussi et la force de ce lien est qualifiée de forte, tandis qu'avec les variables SECS, BN, BC, et NN il existe un lien directe modéré.

Pour les variables Uniformity_of_Cell_Size et Uniformity_of_Cell_Shape, il existe un lien direct fort avec les variables CT, SECS, BN, BC. Par contre, entre ces deux variables UCS et UCSH, il y a un lien direct très fort, soit un pourcentage de 82%, ce qui signifie que le

lien est tellement fort que les deux variables sont presque identiques et l'une peut être pratiquement remplacée par l'autre.

La variable Single_Epithelial_Cell_Size a un lien direct modéré avec les variables CT et BN, cependant elle a un lien direct fort avec les variables UCS, UCSH, BC et NN.

La variable Bare_Nuclei possède un lien direct modéré avec les variables CT et SECS, et un lien fort avec les variables UCS, UCSH, BC et NN.

Les variables Bland_Chromatin et Normal_Nucleoli possèdent un lien direct fort avec les variables UCS, UCSH, SECS et BN ainsi qu'entre elles, alors qu'elles ont un lien direct modéré avec la variable CT.

Analyse multivariée		IC 95%	OR	X^2	P
Clump Thickness	1 - 10	]0.16, 0.25[	1.82	8,588	**<0,0001**
Uniformity of Cell Size	1 - 10	]0.76, 1.57[	1.095	2,324	**0,0205**
Single Epithelial Cell Size	1 - 10	]0.77, 1.54[	1.06	2,183	**0,0294**
Bare Nuclei	1 - 10	]0.32, 0.42 [	1.65	13,840	**<0,0001**
Bland Chromatin	1 - 10	]0.026, 0.13 [	1.48	2,888	**0,0040**
Normal Nucleoli	1 - 10	]0.074, 0.17 [	1.39	4,762	**<0,0001**

Table 3. 14 – Tableau représentant les résultats de l'analyse multivariée sur l'échantillon

Le tableau ci-dessous résume les coefficients de notre modèle :

	Coefficient
Clump Thickness	0,03364
Uniformity of Cell Size	0,01628
Uniformity of Cell Shape	0,01682
Single Epithelial Cell Size	0,01205
Bare Nuclei	0,04952
Bland Chromatin	0,01663
Normal Nucleoli	0,01958
Constante -0,2544	

Table 3. 15 – Tableau des valeurs du coefficient du modèle

6. Validation du modèle :

6.1 Le coefficient de détermination R^2 :

Le tableau suivant indique la valeur de R^2 de notre étude :

R^2(McFadd	0,000	0,888		
R^2(Cox an	0,000	0,680	R^2	0,844
R^2(Nagelk	0,000	0,941	R^2 ajusté	0,841

Table 3. 16 – Tableau des valeurs du coefficient de détermination

6.2 La courbe ROC :

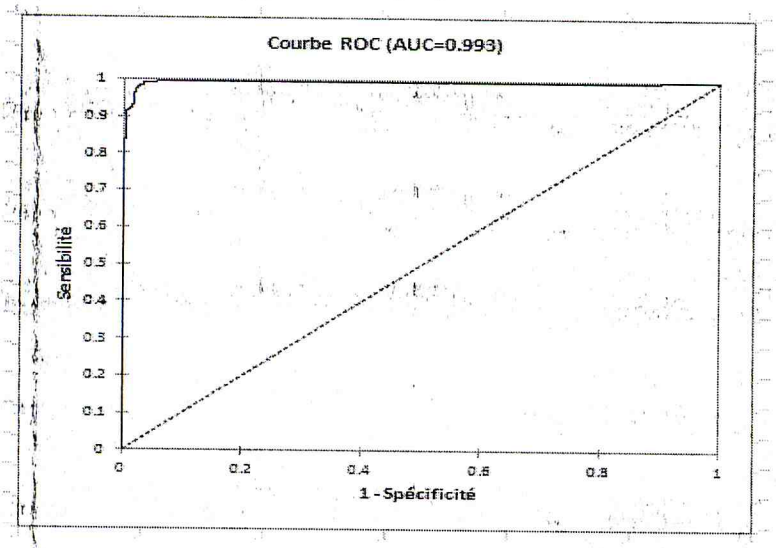


Figure 3. 10 – Tracé représentant les résultats de l'analyse de ROC

Selon la valeur d'AUC, le modèle est qualifié d'excellent car AUC > 0.90.

Conclusion :

Dans ce chapitre, nous avons appliqué l'algorithme de régression logistique par deux fois, la première fois sur des données d'une base de données qu'on a créé nous-même, cette dernière comporte 105 patients du CAC. Tandis que la deuxième fois, les données provenait d'une base de données qu'on a trouvé sur internet et qui concerne une population américaine ou plus particulier de Wisconsin. Le but de faire une deuxième application de l'algorithme était pour prouver que plus grande la base d'apprentissage est, plus fiable sera le modèle et plus précises les résultats de prédiction seront (i.e. on aura plus d'individus classés dans leurs bonne classe que ceux mal classés).

CHAPITRE 4 : REALISATION ET TESTS EXPERIMENTAUX

«J'écoute et j'oublie, je vois et je me souviens,
je fais et je comprends »

Proverbe chinois

1. Introduction :

L'objectif majeur de notre étude est la conception d'un outil de prédiction en employant le pouvoir statistique, dans le but d'automatisation et d'aide à la décision. Il s'agit tout d'abord d'élaborer le modèle qui va être utilisé pour effectuer des prédictions, puis le tester et évaluer ses performances. La plus grande partie du travail était d'arriver à avoir un modèle fiable fournissant des résultats plus ou moins corrects et de programmer cet outil pour qu'il puisse être utilisé par la suite.

2. Application du modèle sur la base de validation :

Maintenant qu'on a obtenu notre modèle à partir de la base de construction et ensuite avoir testé sa fiabilité, on va regarder si notre modèle s'applique bien à la base de validation qu'on a mis de côté auparavant en comparant les résultats que notre modèle nous a retourné avec les valeurs de prédiction déjà existantes.

	Valeur réelles	**Valeurs prédites**
Tumeur bénigne	60	62
Tumeur maligne	40	38

Table 4.1 – Résultats de l'application sur la base de validation

Ces résultats montrent quelques erreurs à prédire la vraie valeur de la variable réponse et c'est probablement due au fait que la base de validation doit aussi être grande ce qui est pas le cas dans notre étude (généralement on choisit les deux bases ½ ½ ou bien 2/3 pour la base d'apprentissage et 1/3 pour celle de validation).

3. Outils de programmation :

Ensuite vient la phase de programmation et de concrétisation du résultat obtenu.

3.1 Le langage PHP : [32]

Les pages web qui circulent sont produites en général à l'aide du code HTML. Ce code sera ensuite interprété au niveau de l'utilisateur par le navigateur.

Le langage PHP (Personal Home Page au début et Hypertext Preprocessor actuellement) est un langage de script coté serveur, et dont le code est directement immergé au milieu du code HTML de la page Web à générer d'une manière dynamique [Chaleat & al 05]. En utilisant PHP sur un serveur web, les pages ou les sites hébergés sur ce serveur deviennent une véritable application web interactive au lieu d'un ensemble de pages statiques plus au moins actualisées.

Pourquoi avons-nous choisi PHP pour programmer notre application ? La réponse à cette question est que nous avons trouvé en ce langage un outil très puissant pour exprimer nos besoins, en voici quelques caractéristiques de ce langage :

- PHP est un produit Open Source tout en étant indépendant des plateformes ;
- PHP a été conçu pour fonctionner sur le web. L'exécution du code est une tâche simple pouvant être accomplie en quelque secondes voire moins, car le moteur de scripts PHP est parfaitement optimisé pour les temps de réponse nécessaires à des applications web ;
- La syntaxe de PHP s'inspire largement du langage C, PHP est un langage multi plateforme. Il a été porté sur de nombreuses stations UNIX, telles que Linux, où il fonctionne aussi bien que sur des machines Windows ;
- PHP offre aux programmeurs plus de 1200 fonctions utilisables dans des applications très variées, il couvre pratiquement tous les domaines en rapport avec le web (fonctions HTTP, accès aux systèmes de fichiers, fonctions de retouche et de génération dynamique des images avec la bibliothèque GD, utilisation des sockets Internet, support de services utilisant les protocoles tel que SNMP, POP3, etc.) ;
- PHP s'interface à de nombreuses bases de données SQL, en offrant des fonctions dédiées pour la prise en charge directe des principaux systèmes de gestion de bases de données relationnelles. PHP prend notamment en charge l'accès au serveur de bases de données MySQL largement utilisé pour les applications web. Et en possédant des fonctions ODBC, PHP permet de s'interfacer d'une manière conviviale à toute base de données ODBC ;

3.2 CSS : [33]

Littéralement *Cascading Style Sheets* (feuilles de style ne cascade), CSS est un langage déclaratif simple pour mettre en forme des pages HTML ou des documents XML. Le langage CSS permet de préciser les caractéristiques visuelles et sonores de présentation d'une page Web : les polices de caractères, les marges et bordures, les couleurs, le positionnement des différents éléments, etc. Le terme de *"Cascading" Style Sheets*sous entend qu'il est possible de définir un style pour une page HTML puis, à l'intérieur de cette même page, de fournir des informations plus précises ou différentes pour présenter certains éléments plus distinctement. Le but de CSS est séparer la structure d'un document HTML et sa présentation.

3.3 Javascript : [34]

Javascript est un langage de script orienté objet principalement utilisé dans les pages HTML. A l'opposé des langages serveurs (qui s'exécutent sur le site), Javascript est exécuté sur l'ordinateur de l'internaute par le navigateur lui-même. Ainsi, ce langage permet une interaction avec l'utilisateur en fonction de ses actions (lors du passage de la souris au-dessus d'un élément, du redimensionnement de la page...). La particularité du JavaScript consiste à créer des petits scripts sur une page HTML dans le but d'ajouter une petite animation ou un effet particulier sur la page. Cela permet en général d'améliorer l'ergonomie ou l'interface utilisateur, mais certains scripts sont peu utile et servent surtout à ajouter un effet esthétique à la page. L'intérêt du JavaScript est d'exécuté un code sans avoir à recharger une nouvelle fois la page.

4. L'interface de l'application :

Nous présentons dans ce qui suit la concrétisation de notre travail. En d'autres termes, réaliser les traitements qu'on a définis dans les chapitres précédents fera l'objet de notre application, et retourner la probabilité que le sujet appartienne à la classe des malades sera son résultat attendu.

4.1 Page d'accueil :

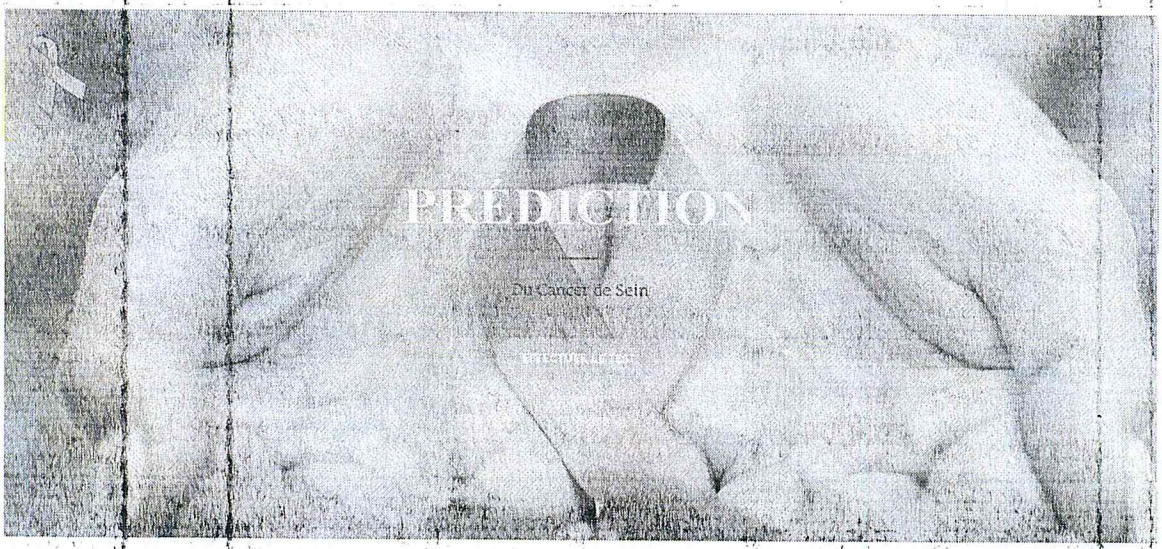


Figure 4. 1 - L'interface de l'application

4.2 Affichage des résultats :

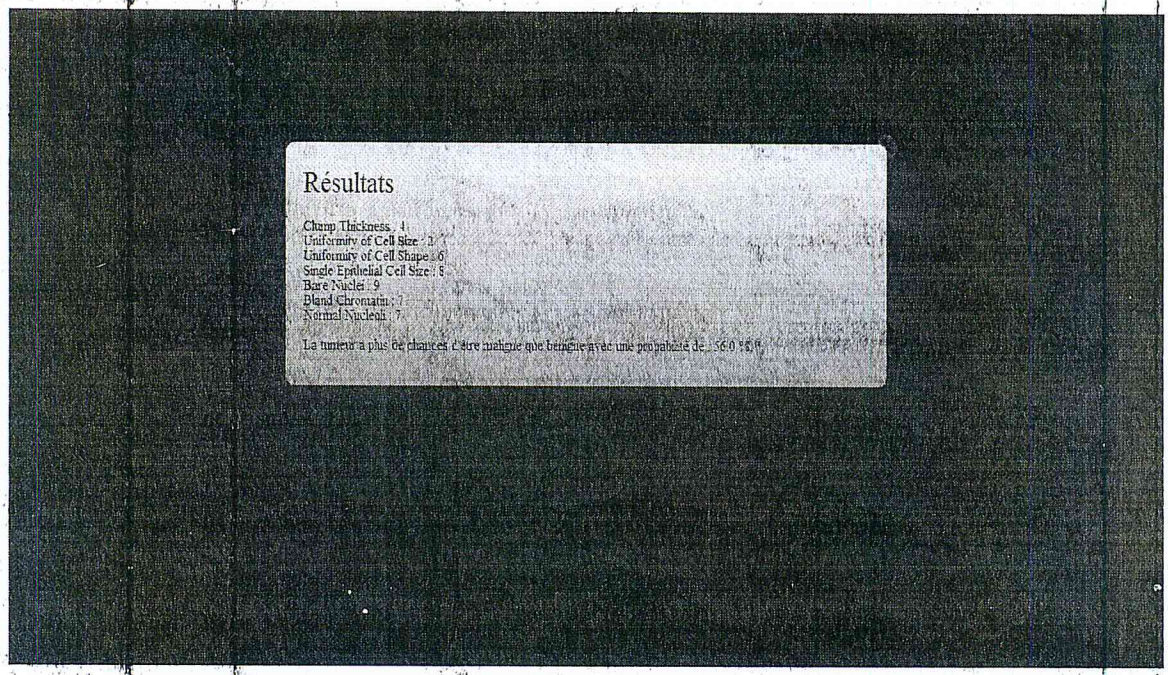


Figure 4. 2 - Affichage des résultats

5. Conclusion :

Nous avons présenté dans ce chapitre l'ensemble des outils de programmation qu'on a utilisé pour réaliser notre instrument de prédiction, ainsi que ses différentes interfaces.

DISCUSSION

«When there are many meanings in a network, you can turn things around in your mind and look at them from different perspectives; when you get stuck, you can try another view. That's what we mean by thinking»

Minsky

Le cancer du sein est la maladie la plus commune et malveillante chez les femmes. Il est souvent difficile de diagnostiquer chez un patient si ce dernier est atteint d'un cancer du sein ou pas, de confirmer sa présence ou déterminer ses caractéristiques (son extension, son agressivité, ….) la chose qui a motivé la recherche dans ce domaine, et l'utilisation d'un outil intelligent spécialisé dans le diagnostic médical a pour ambition de lever cette difficulté avec le temps. La nécessité d'automatiser le diagnostic médical est devenu indispensable, donner plus d'autonomie et d'initiative aux différents modules logiciels spécialisés dans le diagnostic médical poussent à aller plus loin dans la recherche des solutions comme nous avons essayé de faire.

Dans cette étude, nous avons visé l'exploitation du pouvoir de calcul statistique pour prédire l'apparition du cancer du sein. Le défis majeur était de trouver une base de données afin de pouvoir appliquer l'algorithme d'apprentissage machine, la chose qui n'était pas très évidente. Dans la première étude de cas, on a appliqué l'algorithme de régression logistique sur des données provenant d'une population algérienne ou plus précisément du Centre Anti Cancer (CAC) de Blida. L'échantillon d'étude était très restreint tandis que cette méthode donne de bons résultats lorsque l'échantillon est "assez" grand. De plus, la différence entre le nombre de sujets atteints et non atteints était vaste (12 non atteints contre 93 atteints) et c'est due au fait que la notion de dépistage n'est pas très répandue en Algérie. Du coup, et afin de bien pouvoir tester L'algorithme de régression, on l'a appliqué une deuxième fois sur les données d'une base de données publiée sur internet (un benchmark) et cela nous a permis de bien voir la puissance de cette approche.

En étudiant chaque population, nous avons vu que les facteurs de risque diffèrent d'une population à une autre et d'une étude à une autre, sans oublier que ces deux ne partent pas du

même endroit bien que le résultat (avoir un cancer ou pas) soit identique ; dans la première on étudie les facteurs de risque (familiaux etc.) alors que dans la deuxième, on démarre avec un sujet ayant certainement une tumeur et selon ses caractéristique on prédit s'il s'agit d'une tumeur bénigne ou maligne.

Avec la régression logistique réalisée sur l'échantillon d'apprentissage (n = 580), le modèle obtenu contient finalement 7 variables et après validation, ce dernier s'avère être qualifié d'excellent (AUC > 0,9).

Le modèle de régression logistique proposé est une technique paramétrique de modélisation efficace et robuste pour prédire l'apparition du cancer de sein. Le principe mathématique de la méthode est la sélection des variables significatives. Bien que la méthode soit fondée sur une approche probabiliste, sa pertinence a pu être justifiée lors de plusieurs travaux en épidémiologie surtout.

Ce genre d'applications est très utilisé et porte très souvent le nom de « calculateur de risque » (exp : Hearattack risk calculators), mais il reste à préciser qu'on ne peut pas toujours se baser sur ses résultat car dans certains cas, il s'agit de facteurs de risques cachés ou que le système ne prend pas en considération en calculant la probabilité. Aussi dans le cas où le sujet n'est pas de la même population de l'ensemble d'apprentissage, c'est pour ça qu'il est toujours préférable d'indique la population d'étude.

CONCLUSION GENERALE

« Ce n'est pas la fin.
Ce n'est même pas le commencement de la fin.
Mais, c'est peut-être la fin du commencement »

Winston Churchill

L'innovation technologique et le besoin de gérer la mine de renseignements recueillis ont joué un rôle catalyseur dans l'émergence de l'analyse prédictive. En 2006, le magazine informatique *Computerworld* définissait l'*analyse prédictive* comme étant « la branche du forage des données chargée d'établir des probabilités » [20]. Cette définition permet de constater que l'analyse prédictive est un concept unique tourné vers l'avenir et ce, en tentant d'approfondir les connaissances tirées des données rassemblées pour en anticiper la signification et ainsi formuler des prédictions quant à l'avenir.

Les données recueillies constituent l'ingrédient essentiel. L'utilisation de ces données jumelée au désir de pouvoir prédire des événements et aux capacités intelligentes des outils statistiques pourraient être des éléments susceptibles d'amplifier véritablement les conséquences de ces pratiques dans tous les domaines. Après l'assemblage des données, la statistique intervient avec ses algorithmes pour offrir des prédictions les plus précises possibles à l'aide des données disponibles.

Le cadre général de ce mémoire considère la prédiction du risque de la survenue du cancer de sein et ce, en prenant en considération un ensemble de facteurs de risques connus à l'avance.

En premier lieu, nous avons introduit les différents outils et méthodes statistiques dans un cadre théorique général en montrant comment utiliser leurs propriétés pour le traitement des données. Ensuite, pour atteindre notre but, nous avons appliqué la régression logistique qu'est largement utilisée pour l'analyse des enquêtes épidémiologiques sur les données de notre étude. La régression logistique apparaît donc comme un moyen puissant et souple pour analyser la relation entre un ensemble de facteurs et une maladie. L'étude des facteurs de

risque liés à la maladie étudié a été réalisée de manière univariée puis multivariée. Puis, nous avons utilisé les paramètres estimés pour former notre modèle qui, par la suite a été utilisé pour l'appliquer sur des cas réels.

Les résultats obtenus nous paraissent satisfaisants bien que perfectibles, et susceptibles de répondre au moins partiellement à l'objectif fixé.

Et enfin, comme perspectives, il serait intéressant d'utiliser d'autres algorithmes comme la régression de Cox et qui a pour but de crée un modèle de prévision pour les données de la durée à l'événement. Le modèle de Cox génère une fonction de survie qui prévoie la probabilité d'occurrence de l'événement étudié à un instant *t* donné pour les valeurs fournies pour les variables de prédicteur. Il serait intéressant aussi de mettre en réseau l'application sous Internet. Et ce, fournira une meilleure aide à la décision vu comment c'est très utile dans ce domaine bien qu'une machine ne remplacera jamais un médecin.

BIBLIOGRAPHIE

[1] : A. Venot, A. Burgun, C. Quantin, « Informatique médicale, e-Santé : Fondements et applications», Springer-Verlag France, 2013

[2] : H. Larochelle, « Etude de techniques d'apprentissage non supervisé pour l'amélioration de l'entrainement supervisé de modèles connexionnistes », thèse de doctorat en informatique à l'université de Montréal, Décembre 2008.

[3] : L. Candillier, « Contextualisation, visualisation et évaluation en apprentissage non-supervisé », thèse de doctorat en informatique à l'université Charles De Gaulle-Lille 3, Septembre 2006.

[4] : G. Bouchard, « Les modèles génératifs en classification supervisée et applications à la catégorisation d'images et à la fiabilité industrielle », thèse de doctorat à l'université Joseph Fourier-Grenoble1, 2005.

[5] : G. Chêne, M. Savès, «Principaux outils en statistique», Campus Numérique SEME, 28 août 2008.

[6] : Jean Bouyer, Régression logistique - Modélisation des variables quantitatives. Master. Epidémiologie quantitative, Master Recherche Santé Publique. Paris Sud – UVSQ – Paris Descartes - UPEC, 2012.

[7] : El Sanharawi M, Naudet F. Comprendre la régression logistique. J Fr Ophtalmol (2013).

[8] : G. MION, S. HERAULT, N. LIBERT, D. JOURNOIS, «Eléments indispensables de statistiques médicales », URGENCE PRATIQUE - 2010 No102.

[9] : F. X LEJEUNE, « Introduction au logiciel SAS», institut de statistique de l'Université Pierre et Marie Curie Cycle Supérieur 1ère année 2011-12.

[10] : I. Filali, « Détection de la peau humaine dans des images couleurs à l'aide de la régression logistique bayésienne à noyau multinomiale », thèse de magister à l'université de Blida.

[11] : C. Huber, « Bases : Probabilités, Estimation et Tests », Cours de biostatistique I, Université René Descartes.

[12] : The scientific sentence « Mathématiques : Statistiques : Lois de Probabilité : Lois discrètes », Math 2, « http://scientificsentence.net/Equations/Maths2/statistiques ».

[13] : Loi normale, Wikipédia, « https://fr.wikipedia.org/wiki/Loi_normale ».

[14] : R. Choubai, « Les fondements théoriques de la régression logistique et son utilisation en épidémiologie », Université de Sherbrooke, Canada, 2006.

[15] : C. Hurlin, « Econométrie des Variables Qualitatives », thèse de master économétrie et statistique appliquée (ESA) à l'Université d'Orléans.

[16] : L. FOUCAN, « Méthodologie des études épidémiologiques », PAES 2012 – 2013.

[17] : ACSP, « http://www.cpha.ca/fr/programs/portals/substance/prevention», Canada.

[18] : Andre Nkondjock, Parviz Ghadirian, « Facteurs de risque du cancer du sein », MEDECINE/SCIENCES n° 2, vol. 21, février 2005.

[19] : Université Toulouse III paul sabatier, « Methodologie generale de la recherche epidemiologique : les enquetes epidemiologiques », « http://www.medecine.ups-tlse.fr/DCEM2/module1/sous_module1/003_methodologie_generale_CA.pdf »

[20] : Roxane Schaub, Études Epidémiologiques analytiques et biais, « http://www.masterbs.univ-montp2.fr/images/ARC/R%20_SCHAUB_Etudes_epidemiologiques.pdf »

[21] : Delphine Magnin, Philippe Vanhems, Biais et Facteurs de Confusion, « http://www.sf2h.net/SF2H-outils/SF2H_methodo-noso_biais-et-facteurs-de-confusion.pdf »

[22] : José LABARERE, Interprétation d'une enquête épidémiologique : type d'enquête, notion de biais, causalité, « http://www-sante.ujf-grenoble.fr/SANTE/corpus/disciplines/sanpub/methodo/72/leconimprim.pdf »

[23] : S. Benalia, « 40.000 nouveaux cas de cancer par an en Algérie », « http://www.lexpressiondz.com/actualite/160552-40-000-nouveaux-cas-de-cancer-par-an-en-algerie.html »

[24] : Société canadienne du cancer, « Qu'est-ce que le cancer du sein? », «http://www.cancer.ca/fr-ca/cancer-information/cancer-type/breast/breast-cancer»

[25] : Mion G, de Rudnicky S. Eléments indispensables de statistiques médicales – 3 – Lien entre deux paramètres quantitatifs : régression et corrélation. Urgence Pratique 2008 ; 91 : 21-25.

[26] : M. Khaneboubi, « Le test de khi-deux pas à pas », « http://mehdikhaneboubi.free.fr/stat/co/khi_deux.html »

[27] : Wikihow, « Comment calculer la valeur P », « http://fr.wikihow.com/calculer-la-valeur-P »

[28] : Le coefficient de détermination, « http://www.jybaudot.fr/Correlations/coeffdeterm.html »

[29] : Monde MAMBIMONGO WANGO, « Analyse de la vulnérabilité de la santé de la femme: cas du cameroun », Institut Sous-régional de Statistique et d'Economie Appliquée (ISSEA) - 2009

[30] : Wikipédia, « Receiver Operating Characteristic », «https://fr.wikipedia.org/wiki/Receiver_Operating_Characteristic»

[31] : R. Rakotomalala, « Pratique de la Régression Logistique », Université Lumière Lyon 2

[32] : Paluku Vagheni Aloys, « Conception et réalisation d'un système d'inscription en ligne dans les institutions universitaires: Cas de l'UCBC/Beni », Université Chrétienne Bilingue du Congo - 2014

[33] : Jsand Informatique, « Css », « http://www.jsand.net/definition_css.wju »

[34] : Futura High-Tech, « Javascript», « http://www.futura-sciences.com/magazines/high-tech/infos/dico/d/internet-javascript-509 »

[35] : Monde MAMBIMONGO WANGOU, « Analyse de la vulnérabilité de la santé de la femme: cas du cameroun », http://www.memoireonline.com/06/10/3580/m_Analyse-de-la-vulnerabilite-de-la-sante-de-la-femme-cas-du-cameroun25.html

[36] : M. Dramaix-Wilmet, « Modèles de prédiction, Intérêt de la validation ».