Université Ibn Tofaïl
جامعة ابن طفيل
Faculté des Sciences

DEPARTEMENT D'INFORMATIQUE

MASTER BIG DATA AND CLOUD COMPUTING

APPRENTISSAGE AUTOMATIQUE DANS LE DOMAINE MEDICALE

DETECTION DU CANCER DU SEINS

Sous la supervision de Mme RAFALIA Najat

Réalisé par : JDI Zainab et NAJAR Saad

Année Universitaire : 2022/2023

REMERCIEMENT

Nous tenons à exprimer nos remerciements avec un grand plaisir et un grand respect à notre professeur Mme. Rafalia Najat pour ses conseils, sa disponibilité et ses encouragements qui nous ont permis de réaliser ce travail.

Nous exprimons de même notre gratitude envers tous ceux qui nous ont accordé leur soutien, tant par leur gentillesse que par leur dévouement.

RESUME

Aujourd'hui, le cancer du sein est l'une des maladies les plus courantes qui peut causer certaines complications qui peuvent causer parfois la mort. Donc un besoin urgent d'un outil de pronostic pouvant aider les médecins à détecter la maladie à un stade précoce et à recommander les changements de mode de vie nécessaires pour arrêter la progression de cette maladie. L'apprentissage automatique est un besoin urgent d'aujourd'hui pour éliminer l'effort humain et proposer une automatisation plus élevée avec moins d'erreurs. Dans ce projet, un système de détection et de prédiction de cancer du sein est développé en se basant sur une approche de Machine Learning.

ABSTRACT

Today, breast cancer is one of the most common diseases that can cause certain complications, in sometimes worst-case scenario is death. Thus, there is an urgent need for a prognostic tool that can help doctors detect the disease at an early stage and recommend the necessary lifestyle changes to stop the progression of this disease. Machine learning is an urgent need today to enhance human effort and offer higher automation with fewer errors. In this project, a breast cancer detection and prediction system are developed based on a machine learning models

TABLE DE FIGURE

Figure 1 : Apprentissage supervisee..... 10
Figure 2 : L'apprentissage semi-supervise 11
Figure 3 : Apprentissage non supervisé 12
Figure 4:Graque de la régression linéaire simple..... 13
Figure 5:Graphe de la régression logique..... 13
Figure 6: Principe de fonctionnement du KNN 14
Figure 7: Graque du SVM..... 15
Figure 8 : Résultats de reconnaissance pour la 10ièine partition de WDBC 18
Figure 9 : Taux d'erreur de classification des SVM..... 19
Figure 10: Dataset 23
Figure 11:Description du datset après le prétraitement..... 24
Figure 12:Partie du code 24
Figure 13: Visualisation de la corrélation 25
Figure 14:Code de 25
Figure 15: Repartition du " dataset"..... 26
Figure 16:Test des modeles..... 26
Figure 17 : tableaux de performances des algorithmes..... 27
Figure 18: Interface graphique 28

TABLE DE MATIERE

Introduction Generale 7

Objectif 8

Chapitre 1 : Application de machine learning dans le domaine medicale 9

1. Introduction 9
2. Definition du machine learning 9
3. Domaine d'applications du machine learning 9
4. Les types d'apprentissage automatique 10
   - 4.1 Apprentissage supervisé 10
   - 4.2 Apprentissage semi-supervisé 11
   - 4.3 Apprentissage non supervisé 11
   - 4.4 Algorithmes d'apprentissage supervise : 12
5. L'apprentissage automatiques en prédiction et détection des maladies 15

Chapitre 2 : Revue de la Littérature. 17

1. Introduction. 17
2. Approches probabilistes et approches statiques. 17

Chapitre 3 : Implémentation 21

1. Introduction 21
2. Présentation des outils et technologies 21
3. traitement des donnees collectees 23
   - 4.5 Ensemble de donnees : 23
   - 4.6 Preparation des donnees : 24
4. entrainement, test et resultat : 26

5.Conclusion 28

Conclusion Generale 29

Bibliographie 30

INTRODUCTION GENERALE

L'utilisation de l'apprentissage automatique dans la prédiction des maladies est devenue de plus en plus courante dans les domaines médicaux et de la santé. L'apprentissage automatique permet de traiter de grandes quantités de données de manière rapide et efficace, ce qui peut aider les professionnels de la santé à identifier les facteurs de risque, à diagnostiquer des maladies et à prédire leur évolution. Cela utilise des techniques précise pour extraire des modèles à partir de données médicales, telles que les données génomiques, les données de patients et les images médicales. Ces modèles peuvent ensuite être utilisés pour prédire la probabilité qu'un patient développe une maladie, identifier les facteurs de risque, diagnostiquer des maladies, prédire l'évolution d'une maladie et aider à sélectionner le meilleur traitement pour un patient particulier.

Par ailleurs, le cancer du sein est une maladie grave qui affecte des millions de femmes dans le monde entier. Il est important de détecter et de diagnostiquer le cancer du sein à un stade précoce afin d'améliorer les chances de guérison et les résultats cliniques pour les patients. L'utilisation de l'apprentissage automatique pour prédire le cancer du sein peut aider les médecins à détecter la maladie à un stade précoce et à sélectionner le traitement le plus approprié pour chaque patient. Les algorithmes d'apprentissage automatique peuvent être utilisés pour extraire des caractéristiques importantes à partir de données médicales et pour développer des modèles de prédiction qui peuvent être utilisés pour prédire la probabilité qu'un patient développe un cancer du sein.

Cependant, il est important de noter que l'utilisation de l'apprentissage automatique dans la prédiction des maladies soulève également des questions éthiques et de confidentialité. Les données médicales sont souvent sensibles et doivent être protégées contre une utilisation abusive ou inappropriée. Il est donc important de mettre en place des mesures de sécurité et de confidentialité pour garantir que les données sont utilisées de manière responsable et éthique.

OBJECTIF

L’objective de ce projet est de réaliser une application de détection du cancer de siens, à l’aide de l’apprentissage automatique, qui donne des résultats utiles et efficaces, Ceci va aider à prédire si un patient a une tumeur maligne ou bien bénigne.

CHAPITRE 1 : APPLICATION DE MACHINE LEARNING DANS LE DOMAINE MEDICALE

1. INTRODUCTION

Le Machine Learning en médecine peut être utilisé pour analyser de grandes quantités de données cliniques, de tests de laboratoire et d'images médicales afin de détecter des modèles qui pourraient aider à diagnostiquer ou à traiter des maladies. Les modèles peuvent être entraînés à partir de données historiques de patients et peuvent être utilisés pour prédire les résultats de traitement, pour recommander des traitements personnalisés et pour aider à la prise de décision clinique.

Dans ce chapitre on va parler sur le fonctionnement et les techniques de Machine Learning dans le domaine médical et en particulier l'utilisation de ML pour la prédiction du cancer de siens.

2. DEFINITION DU MACHINE LEARNING

L'apprentissage automatique (en anglais, "machine Learning") est une branche de l'intelligence artificielle qui se concentre sur l'utilisation d'algorithmes pour permettre aux machines d'apprendre à partir de données et d'améliorer leur performance au fil du temps sans être explicitement programmées. L'objectif de l'apprentissage automatique est de créer des modèles qui peuvent être utilisés pour prédire des résultats futurs ou classer de nouveaux exemples en fonction de modèles existants.

3. DOMAINE D'APPLICATIONS DU MACHINE LEARNING

Les chercheurs en intelligence artificielle visent toujours à programmer des machines capables d'effectuer des tâches qui requièrent de l'intelligence. Cependant, programmer des machines capables de s'adapter à toutes les situations et éventuellement d'évoluer en fonction de nouvelles contraintes est difficile. L'enjeu est de contourner cette difficulté en dotant la machine de capacités d'apprentissage lui permettant de tirer profit de son expérience. C'est pourquoi parallèlement aux recherches sur le raisonnement automatique se sont développées des recherches sur l'apprentissage par les machines en anglais « machine Learning ». Le principal objectif de ses recherches est la résolution automatique des problèmes complexes par la prise de décision sur la base d'observations de ces problèmes. L'utilisation de l'apprentissage

automatique pour les applications biomédicales connaît une augmentation considérable. Ce regain d'intérêt a plusieurs causes. D'une part, l'application réussie des techniques d'apprentissage automatique dans différents domaines D'autre part, le développement le plus récent est l'avènement des dossiers médicaux électroniques.

4. LES TYPES D'APPRENTISSAGE AUTOMATIQUE

4.1 APPRENTISSAGE SUPERVISE

L'apprentissage supervisé est une méthode d'apprentissage automatique dans laquelle un modèle est entraîné à partir d'un ensemble de données d'entraînement étiqueté. Les exemples d'entraînement consistent en des paires d'entrées et de sorties attendues (étiquettes), et l'objectif de l'algorithme est de trouver une fonction qui relie les entrées aux sorties attendues.

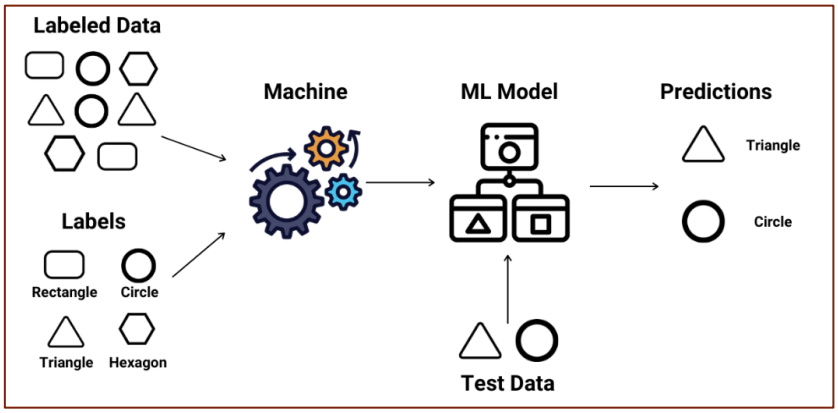

Figure 1 : Apprentissage supervisee

Cette méthode permet de réaliser deux types de tâches

- ***Des tâches de classification***

Ces tâches consistent à attribuer une classe à des objets. Par exemple, dans le milieu bancaire, on peut identifier si une transaction est frauduleuse ou non frauduleuse de manière automatique. On parle de détection d'anomalie. Dans l'industrie, on peut déterminer si oui ou non une machine est susceptible de tomber en panne. On associe une réponse prédéfinie (oui ou non, jaune, rouge, vert ou bleu) à un objet, avant de demander à l'algorithme de réaliser cette classification.

- ***Des tâches de régression***

Ici, on n'attribue pas une classe mais une valeur mathématique : un pourcentage ou une valeur absolue. Par exemple, une probabilité pour une machine de tomber en panne (15 %, 20 %, etc.) ou le prix de vente idéal d'un appartement en fonction de critères comme la surface, le quartier, etc.

4.2 APPRENTISSAGE SEMI-SUPERVISE

L'apprentissage semi-supervisé est une méthode d'apprentissage automatique qui combine des données étiquetées (données pour lesquelles la sortie attendue est connue) et des données non étiquetées (données pour lesquelles la sortie attendue n'est pas connue). Cette technique est souvent utilisée lorsque les données étiquetées sont rares ou coûteuses à obtenir, alors que les données non étiquetées sont plus faciles à collecter en grande quantité. L'objectif est d'utiliser les données non étiquetées pour améliorer les performances de l'algorithme en permettant une meilleure généralisation des données étiquetées.

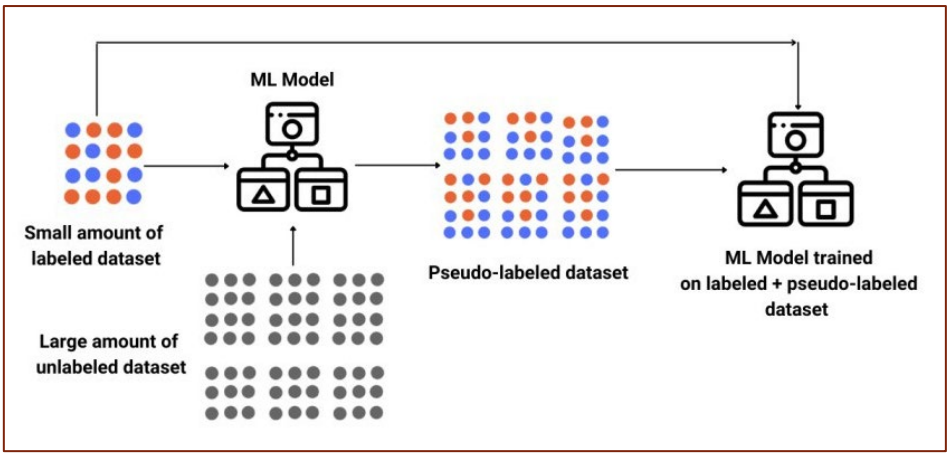

Figure 2 : L'apprentissage semi-supervise

4.3 APPRENTISSAGE NON SUPERVISE

L'apprentissage non supervisé est une méthode d'apprentissage automatique dans laquelle un algorithme explore des données non étiquetées pour en extraire des structures ou des modèles intéressants. Contrairement à l'apprentissage supervisé, il n'y a pas de données étiquetées ou de sortie attendue pour chaque exemple d'entrée. L'objectif est d'explorer les données pour trouver des relations cachées, des groupes similaires ou des tendances intéressantes.

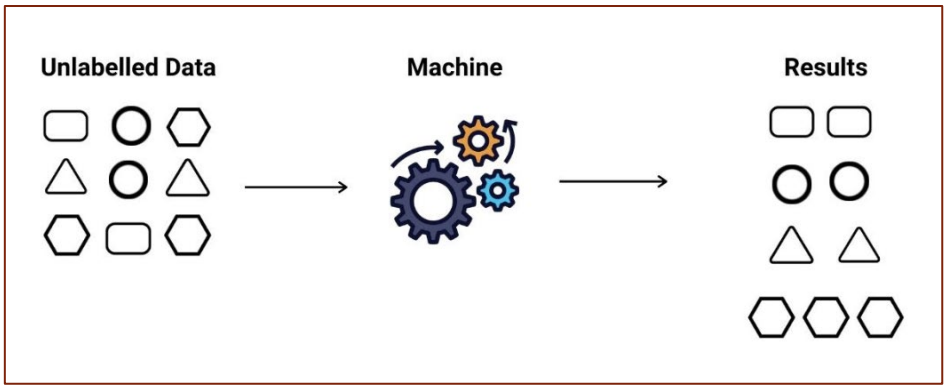

Figure 3 : Apprentissage non supervisé

4.4 ALGORITHMES D'APPRENTISSAGE SUPERVISE :

Afin de créer un apprentissage supervisé, on applique différents algorithmes, selon la méthode employée. Voilà quelque exemple des algorithmes d'apprentissage supervisé

.

a. Régression linéaire :

La régression linéaire est l'un des algorithmes d'apprentissage supervisé les plus populaires. Il est aussi simple et parmi les mieux compris en statistique et en apprentissage automatique. La régression linéaire est un type d'analyse prédictive de base. Le concept général de la régression est d'étudier deux questions :

- Un ensemble de variables prédictives permet-il de prédire une variable de résultat ?

- Quelles sont les variables les plus significatives et ont le plus d'impact sur la variable de résultat ?

On utilise ces estimations de régression pour expliquer les relations entre variable dépendante et une ou plusieurs variables indépendantes. La forme la plus simple de l'équation de régression avec une variable dépendante et une variable indépendante est définie par la formule y = c + b * x, avec y = variable dépendante estimé, c = constante, b = coefficient de régression et x = variable indépendante. On parle ici de Régression linéaire simple. Pour la régression linéaire multiple on écrira y = c + b * x1 +...+ n*xn avec x1 jusqu'à xn les variables indépendantes et b jusqu'à n les coefficients de régression respectifs des variables.

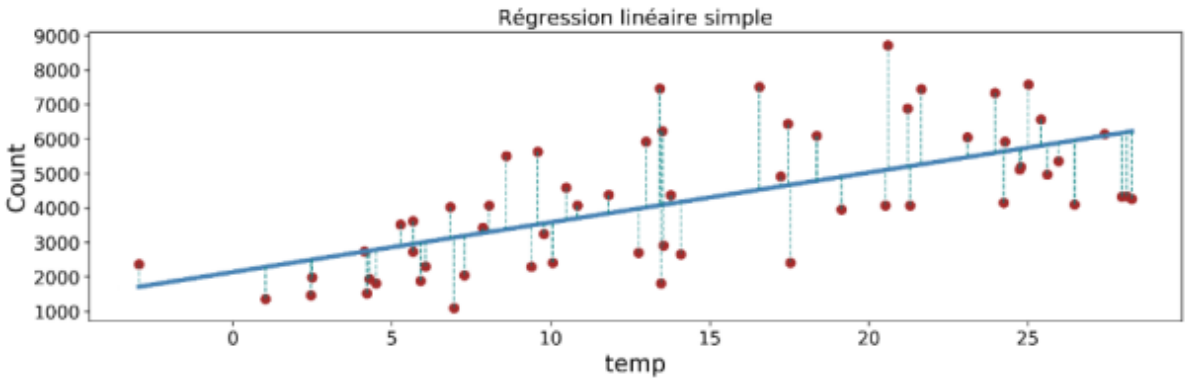

Figure 4:Graque de la régression linéaire simple

b. Régression logistique

Les prédictions de régression linéaire sont des valeurs continues (températures en degrés), Les prévisions de régression logistique sont des valeurs discrètes, c'est-à-dire un ensemble fini de valeurs (Vrai ou faux par exemple). La régression logistique convient mieux à la Classification binaire. Par exemple, on peut considérer un ensemble de données où y = 0 ou 1, où 1 représente la classe par défaut. Pour illustrer on peut imaginer que l'on veuille prédire s'il pleuvra ou non. On aura 1 pour s'il pleut et 0 le cas contraire.

Au contraire de la régression linéaire, la régression logistique, propose le résultat sous forme de probabilités de la classe par défaut. Le résultat appartient donc à l'intervalle [0 :1]. C'est-à-dire qu'il est compris entre 0 et 1, vu qu'il s'agit d'une probabilité. La valeur y de sortie est générée par la transformation de la valeur x, à l'aide de la fonction logistique h (x) = 1 / (1 + e ^ -x). Un seuil est ensuite appliqué pour forcer cette probabilité dans une classification binaire

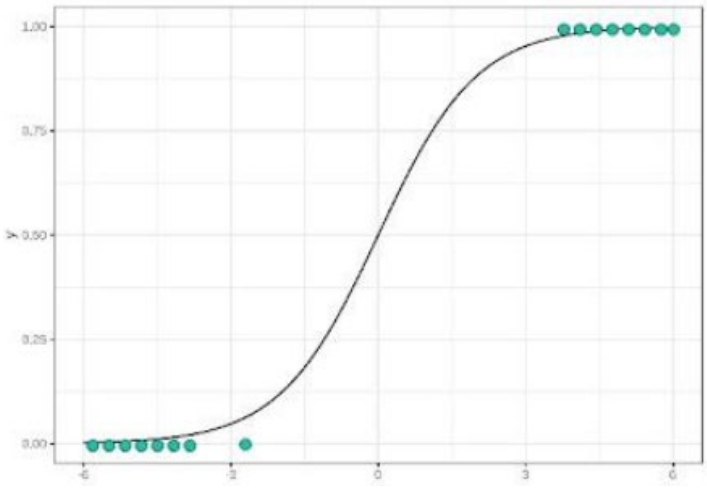

Figure 5:Graphe de la régression logique

c. *K-NN*

L'algorithme K-NN qui signifie k-voisins les plus proches utilise l'intégralité du data set en Tant qu'entraînement, au lieu de diviser se dernier en un training et testing set.
Quand un résultat est requis pour une nouvelle instance de données, l'algorithme KNN parcourt l'intégralité du data set pour rechercher les k-instances les plus proches de la nouvelle Instance ou le nombre k d'instances les plus similaires au nouvel enregistrement, puis renvoie la moyenne des résultats ou le classe à laquelle appartient cette instance si c'est un problème de Classification. L'utilisateur spécifie lui-même la valeur de k.

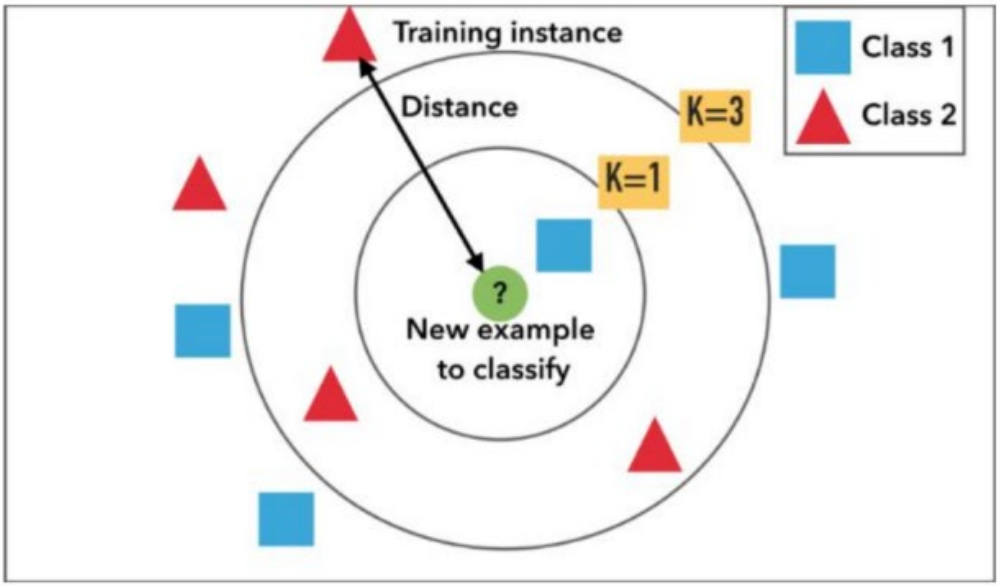

Figure 6: Principe de fonctionnement du KNN

d. *SVM :*

(Support VectorMachine ou Machine à vecteurs de support) : Les SVMs sont une famille d'algorithmes d'apprentissage automatique qui permettent de résoudre des problèmes tant de classification que de régression ou de détection d'anomalie. Ils sont connus pour leurs solides garanties théoriques, leur grande flexibilité ainsi que leur simplicité d'utilisation même sans grande connaissance de data mining.
Comme le montre la figure ci-dessous, leur principe est simple : ils ont pour but de séparer les données en classes à l'aide d'une frontière aussi «simple» que possible, de telle façon que la distance entre les différents groupes de données et la frontière qui les sépare soit maximale.
Cette distance est aussi appelée « marge » et les SVMs sont ainsi qualifiés de « séparateurs à vaste marge », les « vecteurs de support » étant les données les plus proches de la frontière.

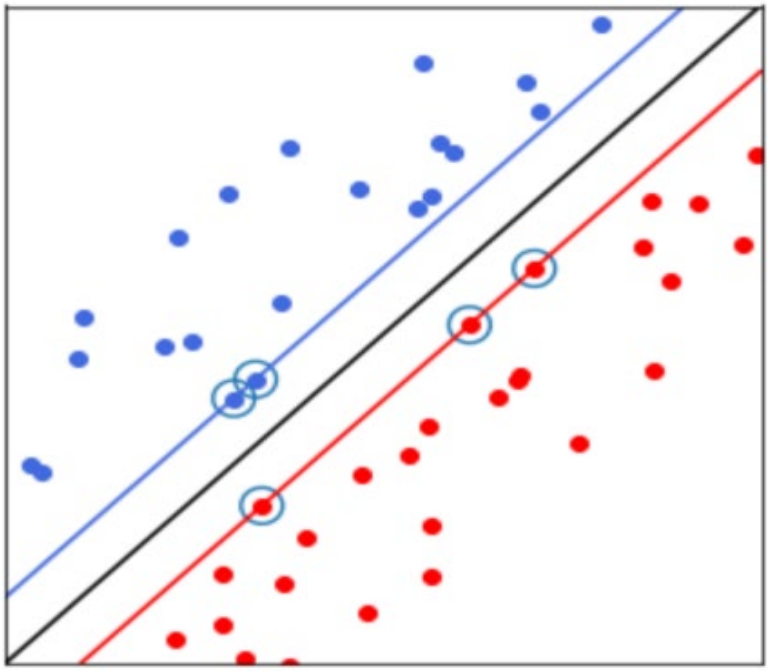

Figure 7: Graphe du SVM

5. L'APPRENTISSAGE AUTOMATIQUES EN PREDICTION ET DETECTION DES MALADIES

La médecine de précision est une nouvelle approche de la recherche clinique et des soins aux patients qui met l'accent sur la compréhension et le traitement des maladies en intégrant les données multimodales ou multimarques d'une personne pour prendre des décisions adaptées au patient. Avec les ensembles de données vastes et complexes générés à l'aide d'approches diagnostiques de médecine de précision, de nouvelles techniques de traitement et de compréhension de ces données complexes étaient nécessaires. Dans le même temps, l'informatique a progressé rapidement pour développer des techniques qui permettent le stockage, le traitement et l'analyse de ces ensembles de données complexes, un exploit que les statistiques traditionnelles et les premières technologies informatiques ne pouvaient pas accomplir.

L'apprentissage automatique, une branche de l'intelligence artificielle, est une méthodologie informatique qui vise à identifier des tendances complexes dans les données qui peuvent être utilisées pour faire des prédictions ou des classifications sur de nouvelles données invisibles ou pour l'analyse de données exploratoires avancées.

L'analyse par apprentissage automatique des données multimodales de la médecine de précision permet d'analyser de vastes ensembles de données et, en fin de compte, de mieux comprendre la santé et les maladies humaines. Cet examen porte sur l'utilisation de l'apprentissage automatique pour les « mégadonnées » de la médecine de précision, dans le contexte de la génétique, de la génomique et au-delà.

Le machine Learning, ainsi que le Deep Learning sont utilisées dans la prédiction des maladies par exemple :

Prédiction de COVID-19 : Le COVID-19 s'est révélé être une maladie virale infectieuse et mortelle, et sa propagation rapide et massive est devenue l'un des plus grands défis du monde.

Les chercheurs ont fourni un examen complet du rôle de l'apprentissage profond et l'apprentissage automatique dans la recherche de techniques de prédiction pour le COVID-19. Un modèle mathématique a été formulé pour analyser et détecter sa menace potentielle. Le modèle proposé est un algorithme de détection intelligent basé sur le cloud utilisant une machine à vecteurs de support (CSDC-SVM) avec des tests de validation croisés. Les résultats expérimentaux ont atteint une précision de 98,4

Le projet SCRUM-Japan Genesis : vise à établir un algorithme, appelé séquen- çage virtuel (VSQ), en utilisant la technologie d'apprentissage profond (DL) et les diagnostics pathologiques pour la prédiction des anomalies du génome du cancer [24].

Prédiction du développement de la maladie d'Alzheimer : pour des patients atteints d'une déficience cognitive légère. [Valenchon, 2019]

Les maladies cardio-vasculaires (MCV) : désignent, pour la plupart, des affec- tions comprenant des veines limitées ou obstruées qui peuvent provoquer une crise cardiaque, une angine de poitrine ou un accident vasculaire cérébral. Le classifica- teur d'apprentissage automatique prédit l'affection en fonction de l'état de l'effet secondaire subi par le patient. [Kumar et al., 2020]

Détection et prédiction prématuré des maladies cardiaques : à l'aide de l'optimisation de KNN qui à donner un résultat de 95.71 % [28].

CHAPITRE 2 : REVUE DE LA LITTERATURE.

1. INTRODUCTION.

Les recherches liées à la détection du cancer du sein se sont multipliées durant la dernière décennie. Beaucoup de travaux se sont dirigés vers la détection de la présence de tissus cancéreux dans le sein et la classification de tumeurs. Les approches utilisées proviennent de plusieurs domaines : probabilités et statistiques, connexionnisme, ainsi que d'autres outils issus de l'intelligence artificielle et des sciences cognitives. À partir de là, une taxonomie des approches récentes de classification dans le cadre du dépistage du cancer du sein a été établie.

2. APPROCHES PROBABILISTES ET APPROCHES STATIQUES.

Dans cette première partie de la revue de littérature, on s'intéresse aux approches probabilistes et statistiques employées en tant que classifieurs pour la détection du cancer du sein. Des méthodes statistiques et probabilistes reviennent dans la littérature, proposant souvent des versions améliorées des approches classiques telles que les réseaux bayésiens et la règle des k plus proches voisins.

Dans (Subhash et al., 2003), les auteurs proposent une approche qui est basée sur une généralisation de la règle des k plus proches voisins pour faire la classification dans le cadre du dépistage du cancer du sein. C'est une méthode qui représente un classifieur non paramétrique mais dont la performance dépend des vecteurs de distributions moyelU1es et des matrices de covariance. Si, de plus, ces distributions sont de nature gaussienne, la performance de cette approche devient intéressante. L'approche a été implémentée et testée sur deux bases de données

portant sur le cancer du sein, WDBC (Wisconsin Diagnosis Breast Cancer) et WBC (Wisconsin Breast Cancer).

La première base de données a été partitionnée et la classification a été effectuée sur chacune

k	k-RNN rule				k-NN rule			
	Confusion matrix		Prob. of false positive false negative	Avg. error rate	Confusion matrix		Prob. of false positive false negative	Avg. error rate
1	335	9	0.026	0.041	337	7	0.020	0.033
	11	128	0.079		9	130	0.065	
2	334	10	0.029	0.039	338	6	0.017	0.048
	9	130	0.065		17	122	0.122	
3	334	10	0.029	0.037	335	9	0.026	0.037
	8	131	0.056		9	130	0.065	
4	334	10	0.029	0.037	338	6	0.017	0.041
	8	131	0.056		14	125	0.101	
5	334	10	0.029	0.037	337	7	0.020	0.035
	8	131	0.056		10	129	0.072	
6	334	10	0.029	0.039	338	6	0.017	0.041
	9	130	0.065		14	125	0.101	

Figure 8 : Résultats de reconnaissance pour la 10ièine partition de WDBC

de ses partitions. La table 1 illustre un exemple de résultat de classification de la dixième partition de cette série de données.

Les résultats obtenus à travers les expérimentations ont été mis en avant par rapport à ceux obtenus en utilisant la règle conventionnelle des k plus proches voisins.

À travers la classification des différentes partitions, les auteurs dans (Subhash et al., 2003) montrent que la méthode utilisée fait preuve de robustesse et offre une meilleure performance que la règle classique des k plus proches voisins. Le meilleur taux de reconnaissance obtenu est de 98.1 % pour la base de données WDBC et de 97% avec WBC.

On note que la méthode proposée dans (Subhash et al., 2003) représente un classifieur à la fois simple à implémenter et d'application générale car non paramétrique, mais dont la performance dépend des vecteurs de distributions moyennes et des matrices de covariance. Par ailleurs, les résultats de classification obtenus sont intéressants dans la mesure où il s'agit d'une classification binaire (classe bénigne/classe maligne). Or, dans les problèmes réels relatifs au cancer du sein, on a souvent affaire à un plus grand nombre de classes de tumeurs.

Dans (Fei et al., 2003), les auteurs proposent une méthode de séparation à vastes marges pour effectuer la classification dans le cadre du dépistage du cancer du sein, soit, une machine à

vecteurs de support (SVM pour Support vector machine). L'approche utilise des mises à jour multiplicatives (<< multiplicative updates ») pour résoudre le problème de programmation quadratique non-négative dans les machines à vecteurs de support. Pour la mise en œuvre des SVMs, les auteurs ont pris en considération un cas simple où, en projetant les données d'entrée dans un espace à grande dimension, les classes deviennent linéairement séparables et l'hyperplan de séparation est contraint de traverser l'origine. Alors, l'hyperplan à marge maximale entre les classes est obtenu en minimisant la fonction de perte :

$$L(\alpha) = -\sum_i \alpha_i + \frac{1}{2}\sum_{ij} \alpha_i \alpha_j y_i y_j \, K(x_i, x_j)$$

où:

{(X;, Yj)} ~I représente les paires d'entrées-sorties à deux classes telles que Yi = ± l, K(Xi, Xj) est une fonction noyau appliquée aux entrées et {ai} ~I représente des valeurs de mise-à-jour multiplicatives.

En respectant la contrainte de non-négativité ai Z O. Le vecteur de l'hyperplan à marge maximal est w = Iiai'YiXi et satisfait les contraintes de marge Yi K(w, Xi) Z 1 pour toutes les données d'apprentissage; a' dénote la localisation du minimum de la fonction de perte.

Kernel	Polynomial		Radial		
Data	$k=4$	$k=6$	$\sigma=0.3$	$\sigma=1.0$	$\sigma=3.0$
Sonar	9.6%	9.6%	7.6%	6.7%	10.6%
Breast cancer	5.1%	3.6%	4.4%	4.4%	4.4%

Figure 9 : Taux d'erreur de classification des SVM

Les résultats de classification sont illustrés dans la table 2. Les expérimentations ont été faites avec une fonction noyau polynomiale et une fonction à base radiale. Les auteurs ont attribué à la fonction polynomiale les degrés k=4 et k=6, et pour la fonction à base radiale des variances α de 1.0 et de 3.0. Les coefficients α_i ont été initialisés à une valeur de 1 d'une manière uniforme dans chaque expérimentation. Les taux d'erreur à la classification obtenus pour la base de données « cancer du sein » varient entre 3.6% et 5.1 %.

Il est à noter que l'approche proposée dans (Fei et al., 2003) aide à pallier aux problèmes relatifs à la programmation quadratique non-négative dans les machines à 8 vecteurs de support. Selon les taux de rappel obtenus à la classification, la performance des SVMs en utilisant cette approche s'avère intéressante. Toutefois, pour la conception de la méthode, les auteurs ont pris en considération un cas simple où, dans un espace à grande dimension, les classes sont linéairement séparables et l'hyperplan séparateur est contraint de traverser l'origine. Cela entraîne une simplification du problème de départ et permet de remettre en question l'applicabilité de l'approche proposée dans des problèmes réels de cas de cancer du sein. Les classes de tumeurs ne sont effectivement pas toujours linéairement séparables. La machine à vecteur de support dans (Fei et al., 2003) constitue un apprentissage hors-ligne. D'autres travaux comprennent des c1assifieurs à vastes marges, tel celoi proposé dans (Baback et Shakhnarovich, 2002), où il est question d'une approche qui utilise des fonctions noyau dyadiques stimulées, ceci étant dans un but de maximiser la marge en offrant la possibilité d'un apprentissage en-ligne.

Dans un autre travail (Huang 2004), les auteurs ont construit un classifieur à partir de la méthode de machines à probabilités Minimax (MPM pour Minimax Probability Machine). Cette méthode permet de générer un effet pire-cas sur la probabilité de mauvaise classification des données, en se basant sur des estimations fiables des matrices de moyennes et de covariance des classes d'apprentissage. En ce qui concerne les MPMs classiques, on y suppose de manière générale que le poids de chaque classe est non biaisé. Or, dans le cas de l'approche proposée par les auteurs dans (Huang 2004), cette hypothèse est ignorée et la méthode est appelée en conséquence : machine à probabilité Minimax biaisée (BMPM pour Biased Minimax Probability Machine). Le modèle a été transformé secondairement en une sorte de problème pseudo-concave avec un minimum local qui est aussi un maximum global. De là, le c1assifieur conçu a été donc applicable à des tâches de classification biaisées et a été évalué avec des données pour le diagnostic du cancer du sein.

CHAPITRE 3 : IMPLEMENTATION

1. INTRODUCTION

L'objectif de ce chapitre est de présenter les étapes de l'implémentation de notre application de détection du siens.

D'abord, on va commencer par la présentation de notre application. Les ressources utilisées dans la création de l'application, et l'interface graphique. Ensuite, le traitement des données collectées, et les résultats obtenus.

Ce chapitre est composé de trois parties à savoir : la présentation outils de développement, le traitement des données collectées, et les résultats obtenus.

2. PRESENTATION DES OUTILS ET TECHNOLOGIES

1. Python :

Python est un langage de programmation simple mais puissant avec d'excellentes fonctionnalités pour le traitement des données linguistiques. Python peut être téléchargé gratuitement sur : http://www.python.org/.

Nous avons choisi Python parce qu'il a une courbe d'apprentissage superficielle, sa syntaxe et sa sémantique sont transparentes, et il a une bonne fonctionnalité de gestion des chaînes. En tant que langage interprété, Python facilite l'exploration interactive. En tant que langage orienté objet, Python permet d'encapsuler et de réutiliser facilement les données et la méthode. En tant que langage dynamique, Python permet d'ajouter des attributs à des objets à la volée et de taper dynamiquement une variable, ce qui facilite le développement rapide.

Python est livré avec une vaste bibliothèque standard, y compris des composants pour la programmation graphique, le traitement numérique et la connectivité. Python est très utilisé dans l'industrie, la recherche scientifique et l'éducation dans le monde entier. Python est souvent loué pour la façon dont il facilite la productivité, la qualité et la maintenabilité des logiciels [w10].

2. Anaconda :

Anaconda est une distribution libre et open source des langages de programmation Python et R appliqué au développement d'applications dédiées à la science des données et à l'apprentissage automatique (traitement de données à grande échelle, analyse prédictive, calcul scientifique), qui vise à simplifier la gestion des paquets et de déploiement. Les versions de paquetages sont gérées par le système de gestion de paquets conda.

3. Pandas :

Pandas est une bibliothèque écrite pour le langage de programmation Python elle nous permet de manipuler et analysé les données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques et de séries temporelles [w12].

4. Scikit-learn :

Scikit-learn (anciennement scikits. Learn) et également connu sous le nom de sklearn) est une bibliothèque d'apprentissage automatique pour le langage de programmation Python.

Elle comporte des divers algorithmes de classification, de régression et de clustering, notamment les machines vectorielles de support, les forêts aléatoires, l'amplification de gradient, kmeans et DBSCAN, et est conçu pour interagir avec les bibliothèques numériques et scientifiques Python NumPy et SciPy [w13].

5. Flask :

Flask est un micro Framework open-source de développement web en Python. Il est classé comme micro Framework car il est très léger. Flask a pour objectif de garder un noyau simple mais extensible. Il n'intègre pas de système d'authentification, pas de couche d'abstraction de base de données, ni d'outil de validation de formulaires. Cependant, de nombreuses extensions permettent d'ajouter facilement des fonctionnalités. Il est distribué sous licence BSD.

6. Machine Learning algorithmes :

- NB (Naive Bayes) : Naive Bayes Classifier est un algorithme populaire en Machine Learning. C'est un algorithme du Supervised Learning utilisé pour la classification. Il est particulièrement utile pour les problématiques de classification de texte.

- SVM (Support Vector Machine) : Les machines à vecteurs de support, ou support vector machine (SVM), sont des modèles de machine learning supervisés centrés sur la résolution de problèmes de discrimination et de régression mathématiques.
- KNN (K-Nearest Neighbor) : un algorithme de Machine Learning qui appartient à la classe des algorithmes d'apprentissage supervisé simple et facile à mettre en œuvre qui peut être utilisé pour résoudre les problèmes de classification et de régression. Dans cet article, nous allons revenir sur la définition de cet algorithme, son fonctionnement ainsi qu'une application directe en programmation.
- LR (Logistic Regression) : La régression logistique est un modèle statistique permettant d'étudier les relations entre un ensemble de variables qualitatives Xi et une variable qualitative Y. Il s'agit d'un modèle linéaire généralisé utilisant une fonction logistique comme fonction de lien.

Traitement des donnees

3. TRAITEMENT DES DONNEES COLLECTEES

4.5 ENSEMBLE DE DONNEES :

	id	clump_thickness	uniform_cell_size	uniform_cell_shape	marginal_adhesion	single_epithelial_size	bare_nuclei	bland_chromatin	normal_nucleoli	mitoses	class
0	1000025	5	1	1	1	2	1	3	1	1	2
1	1002945	5	4	4	5	7	10	3	2	1	2
2	1015425	3	1	1	1	2	2	3	1	1	2
3	1016277	6	8	8	1	3	4	3	7	1	2
4	1017023	4	1	1	3	2	1	3	1	1	2

Figure 10: Dataset

L'ensemble de données contient certains paramètres représentant les caractéristiques et des mesures décrivant la tumeur détecter grâce auxquels le patient peut être identifié avec une tumeur maline ou bénigne. Le « dataset » est composé d'un total de 699 lignes, et 11 colonnes. Les 8 attributs les plus influents qui ont contribué à la prédiction si la tumeur traiter est bénigne ou maline sont.clump_thickness

uniform_cell_size,uniform_cell_shapemarginal_adhesion ,single_epithelial_size ,bare_nuclei, bland_chromatin ,normal_nucleoli.

4.6 PREPARATION DES DONNEES :

1 . Remplir les valeurs manquantes :

De manière générale, les bases de données doivent subir un prétraitement afin d'être adaptées aux entrées du modelé d'apprentissage automatique. Un prétraitement courant consiste à remplir les valeurs manquantes au sein de notre « dataset » ; ainsi que d'éliminer les attribuent qui ne parvienne pas lors de l'entrainement de notre modèle.

```
df4['bare_nuclei']=df4['bare_nuclei'].astype('int64')
moy4=df4['bare_nuclei'].mean()
moy4
```

```
7.627615062761507
```

```
#
df.loc[(df['class']== 4) & (df['bare_nuclei'] =='?'), 'bare_nuclei'] = moy4
```

Figure 12:Partie du code

```
clump_thickness              0
uniform_cell_size            0
uniform_cell_shape           0
marginal_adhesion            0
single_epithelial_size       0
bare_nuclei                  0
bland_chromatin              0
normal_nucleoli              0
mitoses                      0
class                        0
dtype: int64
```

Figure 11:Description du datset après le prétraitement.

Remplir les valeurs manquantes de l'attribut « bare_nuclei » , avec la moyenne de cette dernière.

Apres cette étape , notre enjeu de donnes maintenant ne contient aucune valeurs manquantes ou nulles (Figure 8) .

2.Corrélation :

Une corrélation est une opération pour déterminer s'il existe un lien statistique linéaire fort entre deux variables quantitatives à l'échelle relative (qui identifie les intervalles entre les données et où la position de zéro est définie arbitrairement, comme sur un thermomètre) ou à l'échelle absolue (soit une échelle relative où zéro a une position

absolue : cela permet de quantifier la différence entre deux éléments, comme avec l'échelle de Kelvin).

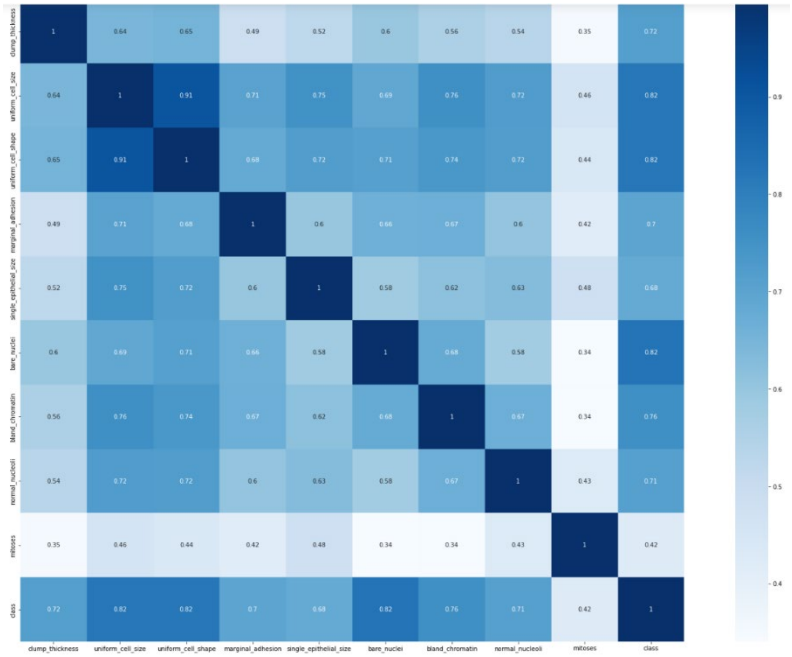

Figure 13: Visualisation de la corrélation

Le calcul de corrélation nous a permis d'identifier la variable qui a une corrélation faible avec l'attribut « class ».

```
cor_target = abs(cor["class"])
irrelevant_features = cor_target[cor_target<0.5] #%50
irrelevant_features
```

```
mitoses    0.42317
Name: class, dtype: float64
```

```
del df['mitoses'] #%42
df.info()
```

Figure 14:Code de

Ce code nous a permis d'extraire l'attribut « mitoses » qui a une corrélation avec « class » moins de 0.5, du coup on a opté pour la suppression de ce dernier vu qui n'affectera pas les résultats de notre traitement dans les étapes suivantes.

4. ENTRAINEMENT, TEST ET RESULTAT :

1.Entrainement et test du modèle :

Par la suite, on va diviser les données en deux partie, données d'apprentissage (training) et données de test afin de tester les modèles en utilisant le code suivant

```
Y=df['class'].values
X=df.drop('class',axis=1).values
X_train, X_test, Y_train, Y_test = train_test_split (X, Y, test_size = 0.25, random_state=1)
```

Figure 15: Repartition du " dataset"

On a diviser notre jeu de données en deux , partie d'entrainement(75 du « dataset ») et partie du test (25 du « dataset »).

```
for name, model in models:
    model.fit(X_train, Y_train)
    predictions = model.predict(X_test)
    print("\nModel:",name)
    print("Accuracy score:",accuracy_score(Y_test, predictions))
    print("Classification report:\n",classification_report(Y_test, predictions))
```

Figure 16:Test des modeles

2 .Choix du modèle :

A cette étape, nous avons effectué une comparaison entre les résultats des algorithmes que nous avons choisie SVM ,KNN ,NB et LR. Les métriques d'évaluation utilisées étaient : l'accuracy, la précision, le rappel et le F1 score. Les formules de calcul de la précision, le rappel et le F1 score sont présentées ci-dessous respectivement dans les équations (1, 2,3 et 4)

• 1 : Précision=TP/(TP+FP)

• 2 : Rappel =TP/(TP+FN)

• 3 : F1score=2 (précision.rappel)/(précision+rappel)=TP/(TP+1/2(FP+FN))

• 4 : accuracy=(TP+TN) /(TP+TN+FP+FN)

telle que :

• TP=nombre de vrai positif

• FN=nombre des faux négatifs

• TN= nombre de vrais négatifs

• FP= nombre de faux positif

Model	Accuracy Score		Classification report			
			Précision	Recall	F1-score	Support
NB	0.97	Bénigne	0.98	0.98	0.98	118
		Maline	0.96	0.96	0.96	57
SVM	0.98	Bénigne	0.98	0.99	0.99	118
		Maline	0.98	0.96	0.97	57
KNN	0.98	Bénigne	0.98	0.99	0.99	118
		Maline	0.98	0.96	0.97	57
LR	0.97	Bénigne	0.97	0.99	0.98	118
		Maline	0.98	0.93	0.95	57

Figure 17 : tableaux de performances des algorithmes

Le tableau présente les résultats obtenus. Nous constatons que les résultats obtenus par le SVM sont meilleurs que ceux obtenus par les autres algorithmes

3. L'interface graphiques

Dans le but de prédire si un patient est un patient a une cellule conquéreuse ou pas , nous avons créé une interface qui permet d'introduire les caractéristiques de la tumeur à savoir : l'épaisseur de la taille, la forme … Puis le système retourne le résultat de prédiction en précisant si la tumeur est maligne ou bégnines.

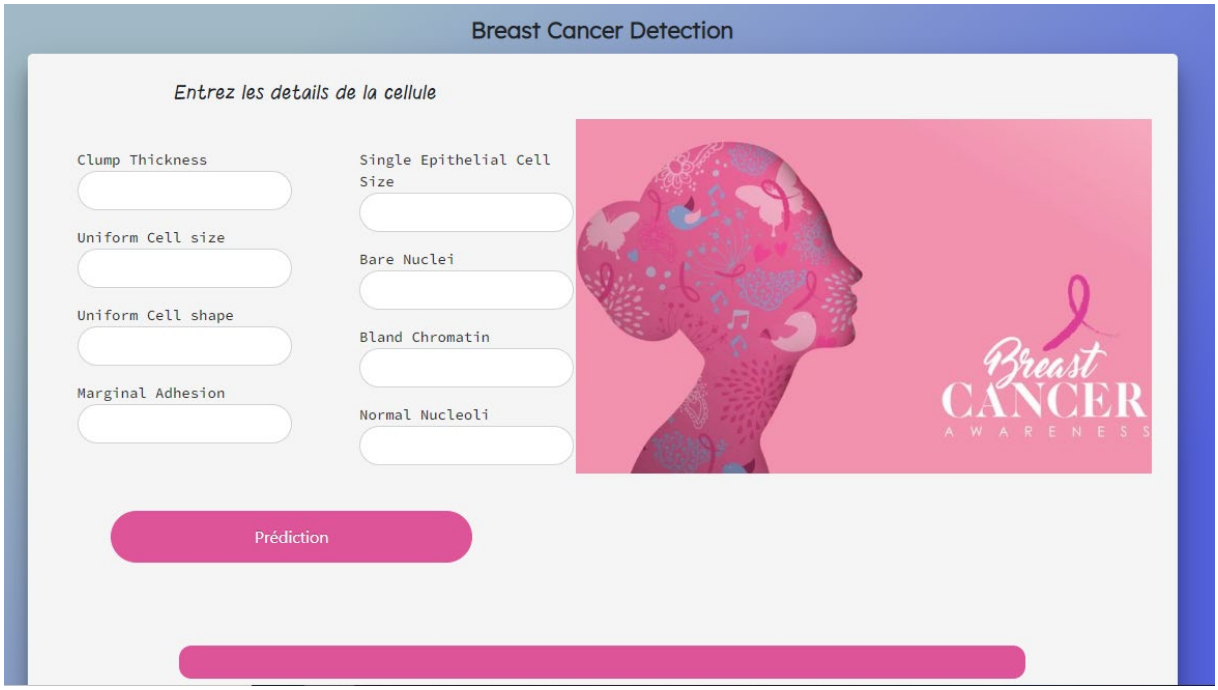

Figure 18: Interface graphique

5.CONCLUSION

Dans ce chapitre, nous avons présenté les différentes étapes que nous avons menées pour parvenir au développement et au bon fonctionnement de notre système de prédiction du cancer de sein.

CONCLUSION GENERALE

La détection du cancer du sein en utilisant le machine Learning est une méthode prometteuse et efficace pour améliorer le diagnostic précoce et la prise en charge de cette maladie. Les algorithmes de machine Learning peuvent être entraînés à reconnaître les caractéristiques des tumeurs malignes et à prédire le risque de développer un cancer du sein.

Les avantages de cette méthode sont nombreux, notamment une amélioration de la précision des diagnostics, une réduction des faux positifs et des faux négatifs, une diminution des coûts associés aux tests de dépistage et une meilleure efficacité de la gestion des soins de santé.

Cependant, il convient de noter que la détection du cancer du sein en utilisant le machine Learning ne doit pas être considérée comme une solution unique. Cette méthode doit être utilisée en complément des méthodes de dépistage existantes, telles que la mammographie et l'examen clinique, et doit être mise en œuvre dans un contexte de soins de santé globaux.

En résumé, l'utilisation d'apprentissage automatique en domaine médicale, surtout en détection du cancer du siens, reste une technique innovante, mais elle doit être utilisée en complément des méthodes existantes et doit être intégrée dans une approche globale des soins de santé.

BIBLIOGRAPHIE

Fisher Igor, et Poland Jan, 2005. « Amplifying the block matrix structure for spectral clustering ». Technical Report, IOSIA, pp. 03-05.

Subhash c., Bagui, Sikha Bagui, Kuhu Pal, et Nikhil R, Pal, 2003. « Breast cancer detection using ranI< nearest neighbor classification rules ». Elsevier Pattern recognition, vol 36, pp. 25-34.

Fei Sha, Lawrence K., Saul, Daniel D., Lee, 2003. « Multiplicative updates for nonnegative quadratic programming in support vector machines ». Advances in Neural Information Processing Systems 15, Sebastian and K. Obermayer, Eds. Cambridge, MA: MIT Press.

Huang Kaizhu, Yang Haiqin, King Irwin, Lyu Michael R, Chan Laiwan, 2004. « Biased minimax probability machine for medical diagnosis ». The 8th International Symposium on Artificial Intelligence and Mathematics, pp. 4-6.

Madden Michael G., 2002. « Evaluation of the Performance of the Markov Blanket Bayesian Classifier Algorithm». CoRR, cs. LG/02l1003.

Esmeir Saher, Markovitch Shaul, 2004. « Lookahead-based Algorithms for anytime induction of decision trees ». 21 th International conference on machine learning, vo169, pp. 33.

De Santo Massimo, Molinara Mario, Tortorella Francesco, Vento Mario, 2003. « Automatic classification of c1ustered microcalsifications by a multiple expert system ». Elsevier Pattern Recognition, 36, pp. 1467-1477.

Karnan M., Thangavel K., Ezhilarasu P., 2008. « Ant colony optimization and a new partic1e swarm optirnizaion algorithm for classification of microcalcifications in mammograms ». The 6th International Conference on Advanced Computing and Communication.

Liu Bo, Abbass Hussein A, McKay Bob, 2004. « Classification mIe discovery with ant colony optimization ». IEEE Computational intelligence bulletin, Vol.3 No. 1.

Parepinelli R. S., Lopes H. S., Freitas A, 2002. « An Ant Colony Algorithm for Classification Rule Discovery ». Data Mining: Heuristic Approach: Idea Group Publishing, H. A a. R. S. a. C. Newton Edition.

Jaganathan P., Thangavel K., Pethalakshmi A, Karnan M., 2007. « Classification mIe discovery wit ant colony optimization and improved quick reduct algorithm ». lAENG International journal of computer science, 33-1, IJCS_33_L9.

Negnevitsky Michael. Artificial Intelligence, 2005. « A Guide to Intelligent Systems ». Addison Wesley, Second Edition.

[w10] magentahtarch./, Dernier accès au site : 06/06/2022

[w11] , magentahttps://research.google.com/colaboratory/faq.html?

hl=fr/, Dernier accès au site : 06/06/2022

[w12] , magentahttps://fr.wikipedia.org/wiki/Pandas/, Dernier accès

au site : 06/06/2022

[w13] , magentahttps://en.wikipedia.org/wiki/Scikit-learn/, Dernier

accès au site : 06/06/2022

[w14] , magentaN.KetkarandE.Santana,Deeplearningwithpython.Springer,

2017,vol.1./, Dernier accès au site : 06/06/2022

[W15] , magentaF.Chollet,\T1\textquotedblleftBuildingautoencodersinkeras,

\T1\textquotedblrightTheKerasBlog,vol.14,2016./, Dernier accès

au site : 06/06/2022