الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التعليم العالي والبحث العلمي

جامعة باجي مختار – عنابـــــــــة

UNIVERSITE BADJI MOKHTAR – ANNABA

BADJI MOKHTAR – ANNABA UNIVERSITY

Faculté : Sciences de l'Ingéniorat
Département : Informatique
Domaine : Mathématique et Informatique
Filière : Informatique
Spécialité : Système Embarqués et mobilité

Mémoire de : MASTER ACADEMIQUE

Thème :

Apprentissage automatique pour la classification du cancer du sein

Présenté par : Arab Chaima

Jury de Soutenance :

Ghanemi	Professeur	Université Badji Mokhtar	Président
Tachouche	Maître de Conférence	Université Badji Mokhtar	Examinateur
Benabbas Farouk	Maître de Conférence	Université Badji Mokhtar	Encadreur

AnnéeUniversitaire : 2020/2021

Remerciements

Nous tenons tout d’abord à remercier Dieu le tout puissant et miséricordieux, qui nous a donné la force et la patience d’accomplir ce Modeste travail.

Nous tenons à remercier tout d’abord notre encadreur : **Mr Benabbas Farouk** pourtous ses conseils, son soutien, son aide, et son orientation.

Nous remercions les membres de jury **Mr Ghanemi** comme président et **Mme Tachouche** comme examinateur d`avoir accepté de juger notre travail.

J’adresse ici mes plus chaleureux remerciements à ma mère, ma sœur, mes amies Hadjer et Zahra.

Nos remerciements s'adressent aussi à mon ami Mohammed et leur cousin Walid pour ses aides précieuses durant mon projet,

Nous remercions tous les enseignants du département d'informatique et des technologies de L’information, aussi notre promotion 2020-2021 Master informatique dans le domaine embarqués.

Nous remercions toutes les personnes qui nous ont soutenues de près ou de loin dans la Réalisation de ce travail.

Dédicaces

Nous dédions notre travail à :

La plus chère au monde, la bougie qui m'ont toujours guidé vers le bon chemin :

Ma mère, la lumière de notre vie, qui a tout fait pour Notre réussite et notre bonheur.

A nos sœurs pour leur aide et courage

À mon ami MOHAMMED,

À Tous ceux qui nous ont aidés à achever ce travail

Table des Matières

Table des Figures

Liste des Tableaux

Résumé

Le cancer du sein est une maladie courante qui touche principalement les femmes, causée par la perturbation de certaines cellules qui se multiplient souvent et forment une masse appelée tumeur. Dans la plupart des cas, le développement du cancer du sein prend plusieurs mois voire plusieurs années. Récemment, des techniques d'apprentissage automatique (ML) ont été utilisées en biomédecine et en informatique pour prendre des décisions en termes de diagnostic et d'analyse afin de lutter contre le cancer du sein. Ce travail de recherche propose un modèle ML pour classer les patients en deux catégories (bénignes ou malignes). Pour ce faire, nous avons comparé les performances entre différents algorithmes d'apprentissage automatique : Support Vector Machine (SVM), Naïve Bayes (NB) et k plus proches voisins (kNN) dans des ensembles de données sur le cancer du sein (WDBC) accessibles au public. L'objectif principal est d'évaluer l'exactitude de la classification des données par rapport à l'efficacité de chaque algorithme en termes d'exactitude, de précision, de sensibilité et de spécificité. Les résultats expérimentaux montrent que SVM donne la plus grande précision (99%).

Abstract

Breast cancer is a common disease that affects mostly women, caused by the disruption of some cells that often multiply and form a mass called a tumor. In most cases, the development of breast cancer takes several months or even years. Recently, machine learning (ML) techniques have been used in biomedicine and informatics to make decisions in terms of diagnosis and analysis in order to combat breast cancer. This research work proposed an ML model to classify patients into two categories (benign or malignant). To achieve this we compared the performance between different machine learning algorithms: Support Vector Machine (SVM), Naïve Bayes (NB) and k-NearestNeighbors (kNN) in publicly available breast cancer (WDBC) datasets.

The main objective is to assess the correctness of the data classification with respect to the effectiveness of each algorithm in terms of accuracy, precision, sensitivity and specificity. Experimental results show that SVM gives the highest accuracy(99%).

الملخص

سرطان الثدي هو مرض شائع يصيب النساء بشكل رئيسي، وينجم عن تمزق بعض الخلايا التي تتكاثر في كثير من الأحيان وتشكل كتلة تسمى الورم. في معظم الحالات، يستغرق تطور سرطان الثدي عدة أشهر أو حتى سنوات. في الآونة الأخيرة، تم استخدام تقنيات التعلم الآلي (ML) في الطب الحيوي وعلوم الكمبيوتر لاتخاذ قرارات تشخيصية وتحليلية في مكافحة سرطان الثدي. اقترح هذا العمل البحثي نموذجًا (ML) لتصنيف المرضى إلى فئتين (حميدة أو خبيثة). للقيام بذلك، قمنا بمقارنة الأداء بين خوارزميات مختلفة للتعلم الآلي: [Support Vector Machine (SVM)] و [Naïve Bayes (NB)] و [k أقرب جيران (kNN)] في مجموعات البيانات الخاصة بسرطان الثدي (WDBC) المتاحة للعامة. الهدف الرئيسي هو تقييم دقة تصنيف البيانات مقابل كفاءة كل خوارزمية من حيث الدقة والدقة والحساسية والخصوصية. أظهرت النتائج التجريبية ن (SVM) يعطي أعلى دقة (99٪).

Introduction Générale

Introduction

Le cancer est l'un des problèmes de santé publique les plus importants au monde, le cancer du sein est le type de cancer le plus courant chez les femmes. Il affecte habituellement les femmes de plus de 50 ans. Cependant, les femmes de tous âges peuvent avoir un cancer du sein et, dans de rares cas, le cancer du sein peut aussi affecter les hommes. Toutes les femmes qui ont entre 50 et 70 ans devraient être régulièrement examinées dans le cadre du programme de dépistage du cancer du sein du NHS (Le National Health Service). Les femmes plus jeunes ayant un risque élevé de cancer du sein peuvent aussi bénéficier du dépistage. La détection des anomalies médicales reste un problème majeur dans l'évaluation des maladies. Le domaine médical profite également des avancées en détection d'anomalies. (Desponds, 2014)

Il est très important que les femmes examinent régulièrement leurs seins pour détecter toute anormalité, et qu'elles fassent examiner tout changement par leur médecin.Dans l'esprit de la plupart des gens, il n'ya pas de diagnostic plus effrayant que le cancer. Le cancer est souvent considéré comme une maladie incurable non diagnostiquée et incroyablement douloureuse. Cette vision effrayante peut être exagérée, il ne fait aucun doute que le cancer est une maladie grave, mais la réalité est que divers cancers peuvent aujourd'hui être traités, éliminés ou ralentis**(Chaumet, 2019)**. Avec les progrès scientifiques en matière de détection et de diagnostic, il y a plus d'espoir que de désespoir dans la plupart des cas de cancer**.**La recherche sur le cancer ne se limite pas aux domaines de la médecine et de la biologie. De nombreux domaines sont concernés par cette recherche, tels que la chimie, la biochimie et la physiologie.

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

Motivation et énoncé du problème

Le diagnostic d'un cancer nécessite la réalisation de plusieurs examen cliniques, biologiques et données pour voir si la personne avait un cancer du sein ou non.

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

Ces dernières années, le domaine de l'intelligence artificielle a amélioré la précision et la rapidité du diagnostic, tout en facilitant la prise de décision clinique et en menant à de meilleurs résultats pour la santé.L'idée générale de ce projet est de combiner une approche ML classique traité par des algorithmes supervisé pour détecter l'anomalie à partir d'une base de données WDBC.

Questions de Recherche

Notre travail consiste à étudier et comparer un ensemble de méthodes d'apprentissage automatique que nous avons découvert pendant notre étude de l'état de l'art dans la communauté de la détection de cancer de sein telle que : LR (Logistic Regression), SVM (Support Vector Machine), NB (Naïve Bayes), K-NN (K-Nearest Neighbors), MLP (Multiplayer perceptron), en utilisant différentes caractéristiques du WDBC pour comparer les méthodes d'apprentissage automatique, suivis d'une discussion des résultats obtenus. Nous tentons donc de faire une évaluation de la performance de plusieurs techniques d'apprentissage machine. Alors le problème quise pose c'est quoi le modèle le plus fréquent parmi les algorithmes d'apprentissage supervisé qui donne un résultat optimal pour la détection ?

Cette étude tente de répondre à ces questions qui circulent dans cette organisation.

Structure du mémoire

Le mémoire est structuré comme suit :

- **Le chapitre 1 :**C’est une présentation des généralités sur le cancer de sein.
- **Le chapitre 2 :**On présente Les méthodes utilisées pour la classification du cancer de sein.
- **Le chapitre 3 :** Description de la base de données utilisée et les techniques de prétraitement que nous avons utilisées suivi par le résultat final.
- Une conclusion et des perspectives futures concluent ce mémoire.

1 Chapitre 1 : Généralités sur le cancer

1.1 Introduction générale

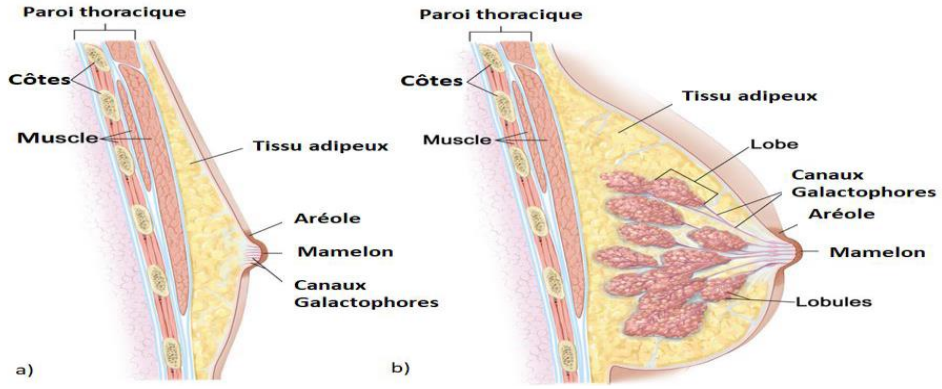
Le Cancer contient un ensemble de maladies qui se caractérisent par la multiplication et la propagation anarchique de cellules anormales. Si les cellules cancéreuses ne sont pas éliminées, l'évolution de la maladie va mener plus ou moins rapidement au décès de la personne touchée.

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

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

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

Les femmes sont plus touchéespar ce cancer que les hommes qui possède aussi cet organe, Les seins chez la femme renferment des glandes mammaires, qui sont responsables de la production du lait maternel pour assurer l'allaitement des nouveau-nés. C'est pourquoi les seins sont des organes accessoires de l'appareil reproducteur féminin.



(Winslow., 2018)

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

Les seins occupent la partie antéro-supérieure du thorax en avant des muscles pectoraux. Ils ne contiennent pas de muscles et sont soutenus par des ligaments.

Le sein correspond à l'organequi appartientà la **glande mammaire (glande exocrine**) qui se développe à partir de la puberté chez la femme.

Le sein est fait de (graisse, tissu conjonctif, glandes et de canaux).

1-La glande mammaire : Est une masse de densité variable, organisée en une vingtaine de lobes. Les glandes produisent du lait quand elles sont stimulées par les hormones de la femme en cours de grossesse.

2- Les lobes : Sont des groupes de glandes qui produisent le lait, Ils sont séparés et maintenus par du tissu « ***conjonctif et adipeux*** ». Chaque sein comporte de 15 à 25 lobules. (Figure1.1)

3-Les ligaments : Sont des bandes serrées de tissu conjonctif qui soutiennent les seins. Ils traversent le sein de la peau jusqu'aux muscles où ils se fixent au thorax.

4-Les canaux : Sont des tubes qui transportent le lait des lobules au mamelon.

5-Le mamelon : Est la région située au centre de l'aréole et d'où sort le lait à une extrémité. Le mamelon est fait de fibres musculaires. Quand ces fibres se contractent, le mamelon durcit, ou pointe vers l'extérieur.

6-L'aréole : Est la surface ronde, rosée ou brunâtre qui entoure le mamelon. Elle contient de petites glandes qui libèrent, ou sécrètent, une substance huileuse qui agit comme lubrifiant pour le mamelon et l'aréole.(Vandenbossche, 2016)

1.3 Les symptômes :

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

- Nodosité dure ou solide dans le sein ou le creux axillaire.
- Modification de la peau : notamment rougeurs ou aspect en peau d'orange.
- Ecoulement du mamelon.
- Rétraction ou aspect bombé (épaississement) de la peau ou du mamelon.

1.4 Les Traitements :

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

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

- Taille et emplacement de la tumeur.
- Taille du sein atteint.
- Propagation du cancer aux ganglions lymphatiques.
- Traitements déjà reçus pour le cancer du sein.
- **Radiothérapie :** Lors de la radiothérapie externe, on a recours à un appareil pour diriger la radiation à travers la peau vers la tumeur et le tissu qui l'entoure.
- **Chimiothérapie :**La chimiothérapie est un traitement courant du cancer du sein. On l'administre souvent après la chirurgie d'un cancer du sein précoce afin de réduire le risque de réapparition de la maladie.
- **Hormonothérapie :**On administre souvent une hormonothérapie pour traiter le cancer du sein dont les récepteurs hormonaux sont positifs. Les femmes ménopausées reçoivent des médicaments hormonaux différents de ceux qu'on administre aux femmes pré-ménopausées.
- **Immunothérapie :**L'immunothérapie aide à renforcer ou à rétablir la capacité du système immunitaire à combattre le cancer. On l'appelle parfois thérapie biologique.On peut administrer une immunothérapie pour :

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

1.5 Les causes et les Facteurs de risques

Quelques facteurs augmentent le risque d'apparition du cancer du sein. Les plus importants ne sont pas modifiables :

Age : Le risque augmente sensiblement après 50 ans, mais la femme plus jeune peut être également touchée

Antécédents familiaux : les femmes dont les sœurs, les mères ou les filles ont un cancer du sein présentent un risque plus important. En particulier si les membres de la famille sont atteints avant l'age de 50 ans.

Prédispositions héréditaires : environ 5 à 10 % de tous les cancers du sein sont déclenchés par une prédisposition héréditaire. Les femmes concernées sont souvent malades avant 50 ans.

Métabolisme hormonal naturel : le risque de cancer du sein est légèrement plus élevé chez les femmes dont la première menstruation s'est produite avant l'âge de 12 ans, dont la dernière a eu lieu après l'âge de 55 ans, chez les femmes qui n'ont pas d'enfants ou qui ont accouché après l'âge de 30 ans. Le mode de vie joue également un rôle. Les facteurs suivants peuvent légèrement augmenter le risque:

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

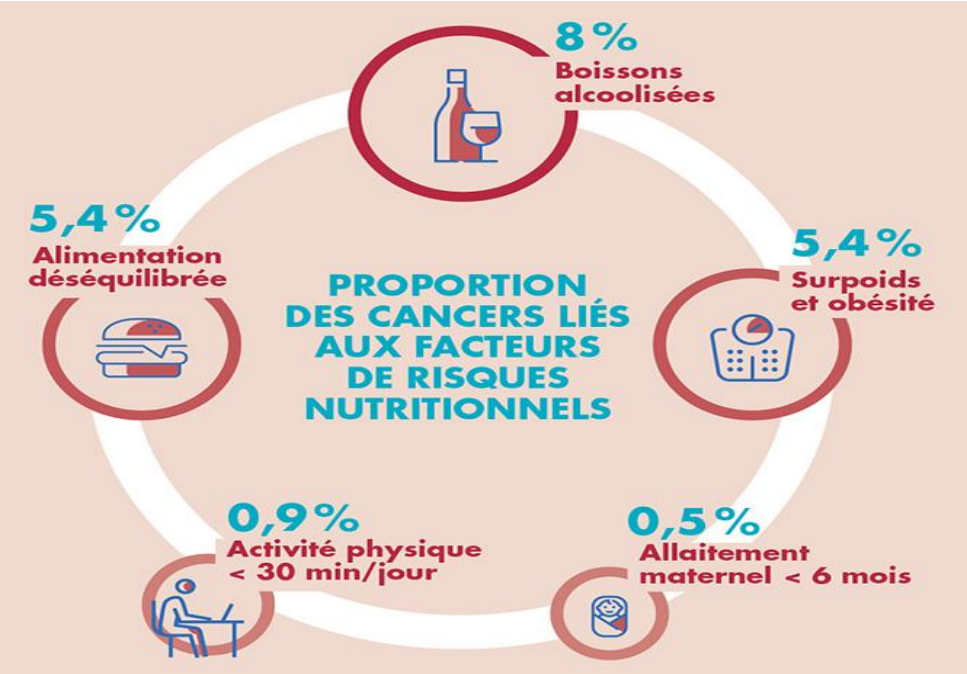


Figure 1-2 : les facteurs de risque du cancer du sein.(wiliam, 2020)

1.6 Tests diagnostiques

Pour diagnostiquer un cancer du sein, on a recours en premier lieu à deux procédés:

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

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

Parfois par une IRM (L'imagerie par résonance magnétique)**(Issy-les-Moulineaux, 2018)**

1.7 Conclusion

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

2 Chapitre 2 : L'état de l'art des méthodes de classification

2.1 Introduction :

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

L'apprentissage machine est le domaine s'applique à comprendre et générer la faculté de l'apprentissage humain d'après d'un système artificiel. Il s'agit, très schématiquement, de concevoir (des algorithmes et des méthodes) permettant d'extraire l'information adéquate de données, ou d'apprendre des comportements à partir d'exemples. Ainsi, le but essentiel de ML est de construire la relation entre les objets et leurs catégories pour la prédiction et la découverte des connaissances **(Mohamed Bouguessa, 2015)**. Nous allons à présent décrire les principaux algorithmes du Machine Learning qui vont nous permettre de réaliser cet apprentissage.Nous distinguons deux partie « l'apprentissage supervisé et de l'apprentissage non supervisé ». Dans ce projet, nous n'aborderons que l'apprentissageMachine supervisé.

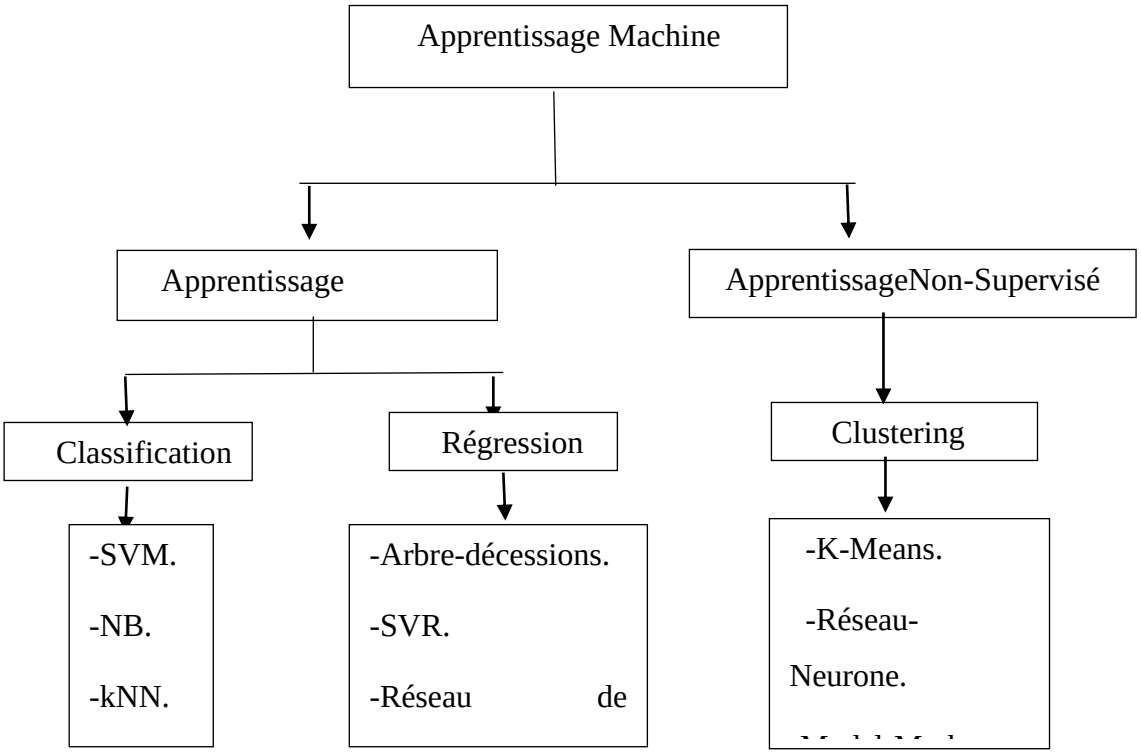


Figure 2-1 : Taxonomie de model d'apprentissage Automatique.

2.2 Les Concepts et les terminologies de base :

L'apprentissage automatique (ou artificiel) (machine-Learning en anglais) est un des champs d'étude de l'intelligence artificielle. Il fait référence à la capacité d'un système à acquérir et intégrer de façon autonome des connaissances **(Karpagavalli, et al., 2009).** L'apprentissage automatique fait référence au développement, l'analyse et l'implémentation de méthodes qui permettent à une machine (au sens large) d'évoluer et de remplir des tâches associées à une intelligence artificielle grâce à un processus d'apprentissage. Cet apprentissage permet d'avoir un système qui s'optimise en fonction de l'environnement, les expériences et les résultats observés **(A. Cornuéjols, 2002).**Les méthodes de l'apprentissage automatique forment une classe de techniques attrayantes pour l'accomplissement des tâches d'extraction de connaissances évoquées. Bien choisis, ces outils peuvent être amenés à accompagner, voire à remplacer l'opérateur humain.

2.3 Pré-traitements des données :

Le prétraitement des données est l'une des tâches d'exploration de données les plus courantes. Il implique un processus d'exploration de données qui prépare les données et les convertit sous une forme appropriée. Le prétraitement des données vise à réduire la taille des données, à trouver des relations entre les données, à normaliser les données, à supprimer les valeurs aberrantes et à extraire des caractéristiques des données. Il comprend plusieurs technologies telles que le nettoyage, l'intégration, la transformation et la réduction des données **(Bhaya, 2017).**

2.4 Normalisation :

La mise à l'échelle des caractéristiques est une technique utilisée pour normaliser la gamme de variables indépendantes ou de caractéristiques de données. Dans le prétraitement des données, également connu sous le nom de normalisation des données, il est généralement utilisé dans l'étape de prétraitement des données**.**

2.5 La Classification :

Nous utilisons l'ensemble de données d'apprentissage pour obtenir de meilleures conditions aux limites qui peuvent être utilisées pour déterminer chaque catégorie cible. Une fois les conditions aux limites déterminées, la tâche suivante consiste à prédire la catégorie cible. L'ensemble du processus est appelé classification. Il existe deux types de classification, supervisée et non supervisée. Dans la classification supervisée, des connaissances prédéfinies disponibles sont nécessaires, tandis que dans la classification non supervisée, parfois appelée analyse de données en cluster ou exploratoire, il n'est pas nécessaire de disposer de données étiquetées prédéfinies **(Agrawal, et al., 2004)**.

2.5.1 Apprentissage Non-supervisé :

Dans le clustering, la catégorie de l'objet est inconnue, Cependant, nous connaissons les règles à classer, et nous connaissons aussi les caractéristiques (variables indépendantes) qui peuvent décrire la classification de l'objet. Il n'y a pas d'exemple de formation pour vérifier si la classification est correcte. Par conséquent, les objets sont simplement affectés à des groupes en fonction de règles données.

2.5.2 Apprentissage Supervisé :

L'apprentissage supervisé est de construire un modèle incisif de la distribution des étiquettes de classe en termes de caractéristiques prédictives. Le classificateur résultant est ensuite utilisé pour attribuer des étiquettes de classe aux instances de test où les valeurs des caractéristiques prédictives sont connues, mais la valeur de l'étiquette de classe est inconnue (Kotsiantis et al., 2006).En autre terme L'apprentissage supervisé a pour but d'établir des règles de comportement àpartir d'une base de données contenant des exemples de cas déjà étiquetés. La base dedonnées est en principe un ensemble de couples entrées / sorties {(X, Y)}. Le but estd'apprendre à prédire pour toute nouvelle entrée X, la sortie Y, (**Bellaterra, 2013)**.Dans l'apprentissage supervisé illustré à la **figure 2.2**, l'apprenant dispose de deux ensembles de données, un ensemble d'apprentissage et un ensemble de test. L'idée est que les apprenants « apprennent » à partir d'un ensemble d'exemples étiquetés dans l'ensemble

d'apprentissage afin qu'ils puissent identifier les exemples non étiquetés dans l'ensemble de test avec la plus grande précision possible **(Learned-Miller, 2014)**

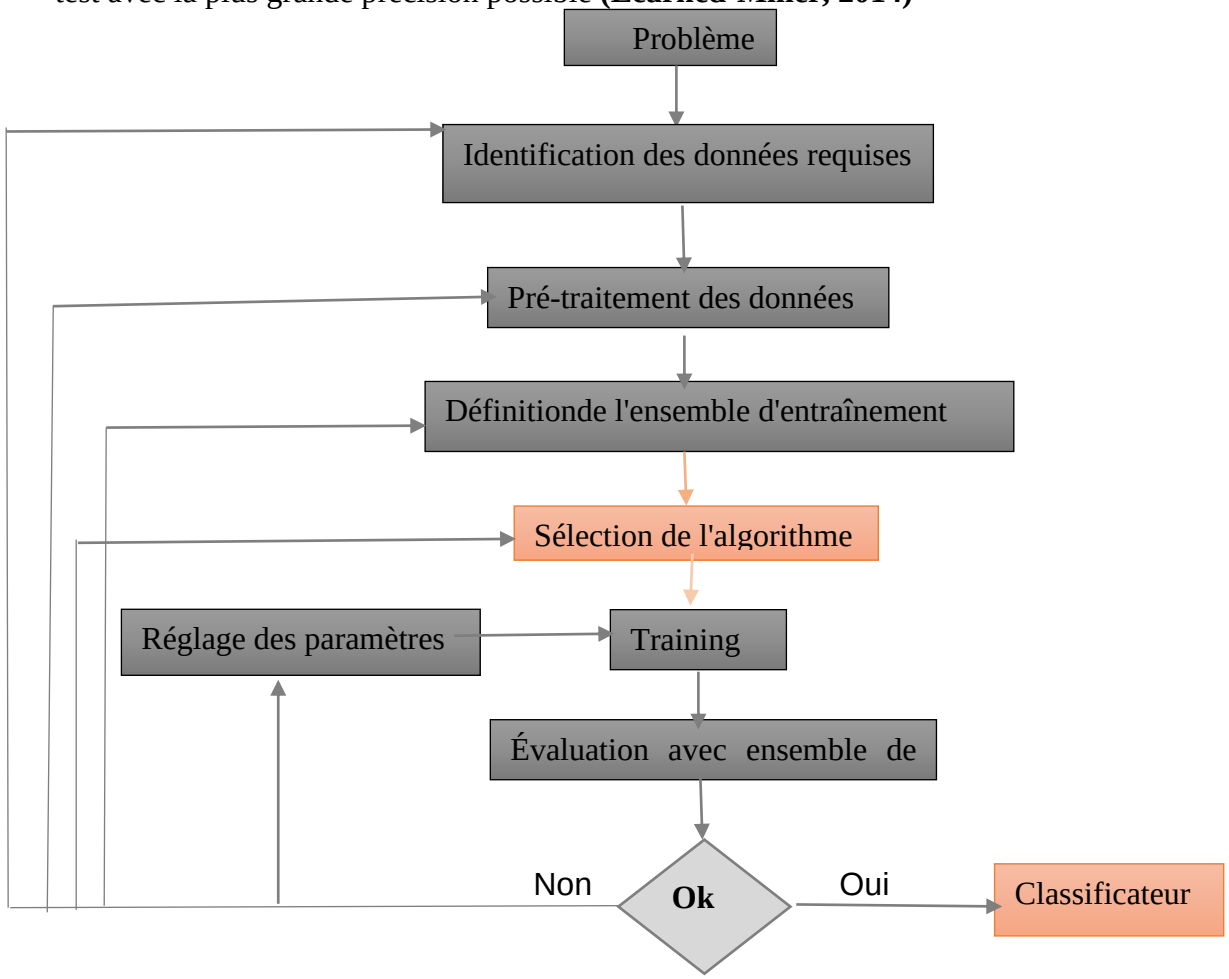


Figure 2-2 : : Les processus d'apprentissage automatique supervisé.

2.5.3 Les modèles d'apprentissage supervisé :

Nous utilisons l'ensemble de données d'apprentissage pour obtenir de meilleures conditions aux limites qui peuvent être utilisées pour déterminer chaque catégorie cible. Une fois les conditions aux limites déterminées, la tâche suivante consiste à prédire la catégorie cible. L'ensemble du processus est appelé classification. Il existe deux types de classification, supervisée et non supervisée. Dans la classification supervisée, des connaissances prédéfinies disponibles sont nécessaires, tandis que dans la classification non supervisée, parfois appelée analyse de données en cluster ou exploratoire, il n'est pas nécessaire de disposer de données étiquetées prédéfinies **(Agrawal, 2004)**.

2.5.3.1 Régression logistique :

L'analyse de régression est largement utilisée pour la prévision, la compréhension des variables indépendantes liées à la variable dépendante et l'exploration de la forme de ces relations. Dans des circonstances limitées, l'analyse de régression peut être utilisée pour déduire la relation causale entre la variable indépendante et la variable dépendante. La fonction logistique est une courbe en forme de S qui peut prendre n'importe quel nombre réel et le mapper en une valeur comprise entre 0 et 1, mais jamais exactement à ces limites. (Jason, 2013)

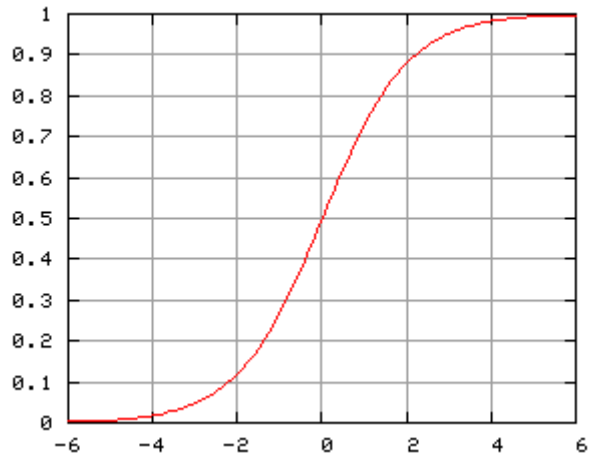


Figure 2-3 : La fonction logistique (sigmoïdes).(Wikipedia., 2005)

2.5.3.2 Naïve Bayes méthode :

La méthode de classification naïve de Bayes est un algorithme d'apprentissage automatique supervisé qui classe un ensemble d'observations selon les règles déterminées par l'algorithme lui-même. L'outil de classification doit d'abord être entraîné sur un ensemble de données d'entraînement, qui affiche la catégorie attendue en fonction de l'entrée. Dans la phase d'apprentissage, l'algorithme développe ses règles de classification sur cet ensemble de données et les applique à la classification de l'ensemble de données prédit dans la deuxième

étape. Le classificateur NB signifie que la classe des données d'apprentissage est connue et fournie. La classification NB a obtenu des résultats remarquables dans de nombreuses applications quotidiennes, ce qui en fait l'algorithme préféré des outils d'apprentissage automatique **(BENZAKI, 2016-2017).**

La base de la classification **NB** est le théorème de Bayes, qui simplifie l'hypothèse d'indépendance entre toutes les paires de variables, qui est appelée naïve **(Sung, 2019)**. Le théorème de Bayes énonce la relation suivante, étant donné la variable de classe y et le vecteur de fonction dépendante x_1 à x_n **(Godé, 2020)**

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

$P(A/B)$: Est la probabilité postérieure de classe (y, cible) donnée prédicteur (x, attributs).

$P(A)$: est la probabilité a priori de classe.

$P(B/A)$: est la vraisemblance qui est la probabilité du prédicteur de la classe donnée.

$P(B)$: est la probabilité a priori du prédicteur.

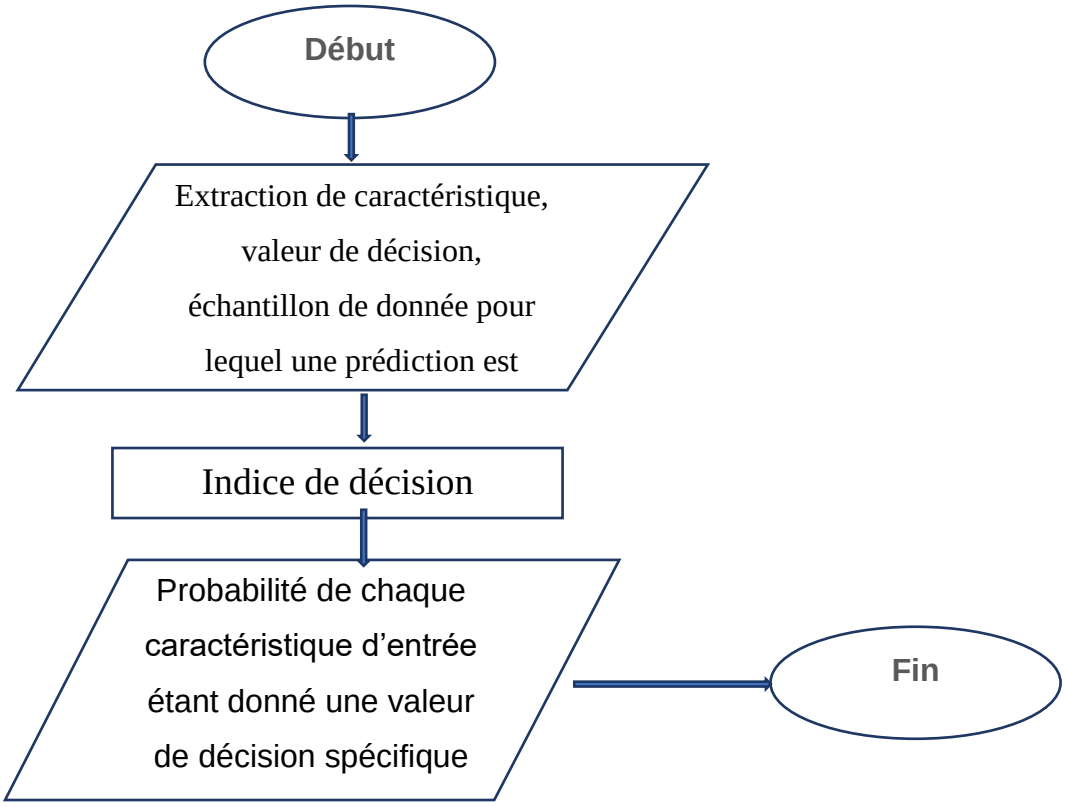


Figure 2-4 : Organigramme de l'algorithme Naïve Bayes.

2.5.3.3 Support Vecteur Machine:

SVM est une technique de classification ML supervisée couramment utilisée dans le domaine du diagnostic et du pronostic du cancer. SVM sépare les catégories en sélectionnant des échantillons clés de toutes les catégories appelés vecteurs de support et en générant une fonction linéaire qui utilise ces vecteurs de support pour les diviser aussi largement que possible **(Dana Bazazeh, 2016)**.

Support Vector Machines (SVM) a été initialement développé par Vapnik et ses collègues dans les années 1960 sur la base de la théorie de l'apprentissage statistique de Vapnik & Chervonenkis en 1992. La SVM a été appliquée avec succès à de nombreuses applications, notamment la reconnaissance de l'écriture manuscrite, la prédiction de séries chronologiques, la reconnaissance vocale, les problèmes de séquence de protéines, le diagnostic du cancer du sein, etc**.**

L'algorithme de SVM repose sur le vecteur de support, qui est l'ensemble de données le plus proche de la limite de décision, ce qui rend SVM différent des autres technologies. En effet, la suppression d'autres points de données plus éloignés de l'hyperplan de décision ne modifie pas la limite comme la suppression des vecteurs de support **(Md. Toukir Ahmed, 2020)**.L'algorithme SVM est un classificateur dit linéaire, ce qui signifie que dans des conditions parfaites, les données doivent être linéairement séparables. Il permet de trouver le meilleur séparateur (ligne, plan ou hyperplan) qui sépare le mieux les deux types

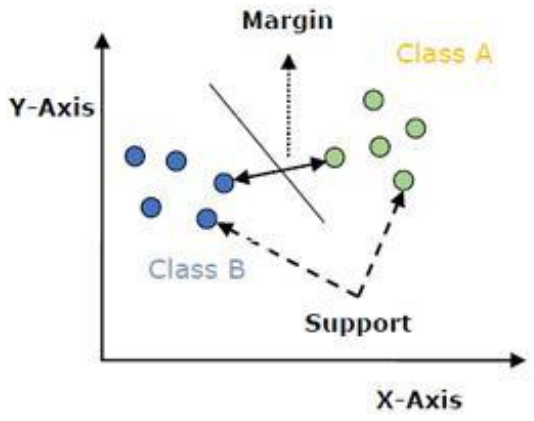


Figure 2-5 : Principe de Classification en SVM (support Vecteur Machine).

2.5.3.4 K-Nearest Neighbors :

Le K Nearest Neighbors est une méthode de classification supervisée intuitive, souvent utilisée en apprentissage automatique. Il s'agit d'une généralisation de la méthode du plus proche voisin (NN). NN est un cas particulier de KNN La méthode K plus proche voisin (KNN) a été largement utilisée dans les applications d'exploration de données et d'apprentissage automatique en raison de sa mise en œuvre simple et de ses performances exceptionnelles. Cependant, il a été prouvé que la définition de toutes les données de test à la même valeur dans la méthode KNN précédente rend ces méthodes peu pratiques dans les applications pratiques.

Supposons qu'il y ait deux catégories, la catégorie A et la catégorie B, et que nous ayons un nouveau point de données x1, donc à laquelle de ces catégories le point de données appartiendra. Afin de résoudre ce genre de problème, nous avons besoin d'un algorithme K-NN. Avec l'aide de K-NN, nous pouvons facilement identifier la catégorie ou la catégorie d'un ensemble de données spécifique. Considérez l'image suivante :

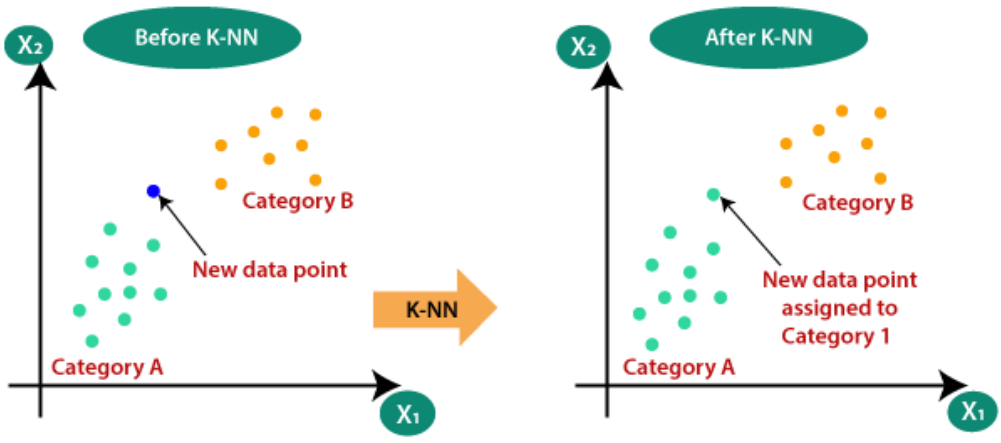


Figure 2-6 : le fonctionnement de K-NN pour un apprentissage automatique.

Le fonctionnement de K-NN peut être expliqué sur la base de l'algorithme suivant :

Étape 1 : Sélectionnez K voisins

Étape 2 : Calculer la distance euclidienne du nombre voisin K

Étape 3 : Selon la distance euclidienne calculée, prenez K voisins les plus proches.

Étape 4 : Parmi les k voisins, comptez le nombre de points de données dans chaque catégorie.

Étape 5 : Attribuez le nouveau point de données à la catégorie avec le plus grand nombre de voisins.

Étape 6 : Notre modèle est prêt.

2.5.3.5 Perceptron multicouche :

Le MLP est basé sur la procédure de supervision illustrée à ***la figure2.7*** *c'est-à-dire que le réseau construit un modèle basé sur des échantillons de données avec production. La relation entre le problème et la solution peut être très générale, par exemple, simuler la richesse des espèces en termes de qualité d'habitat (entrée) ou d'abondance animale (sortie)* **(Brahim, 2016-2017)**.

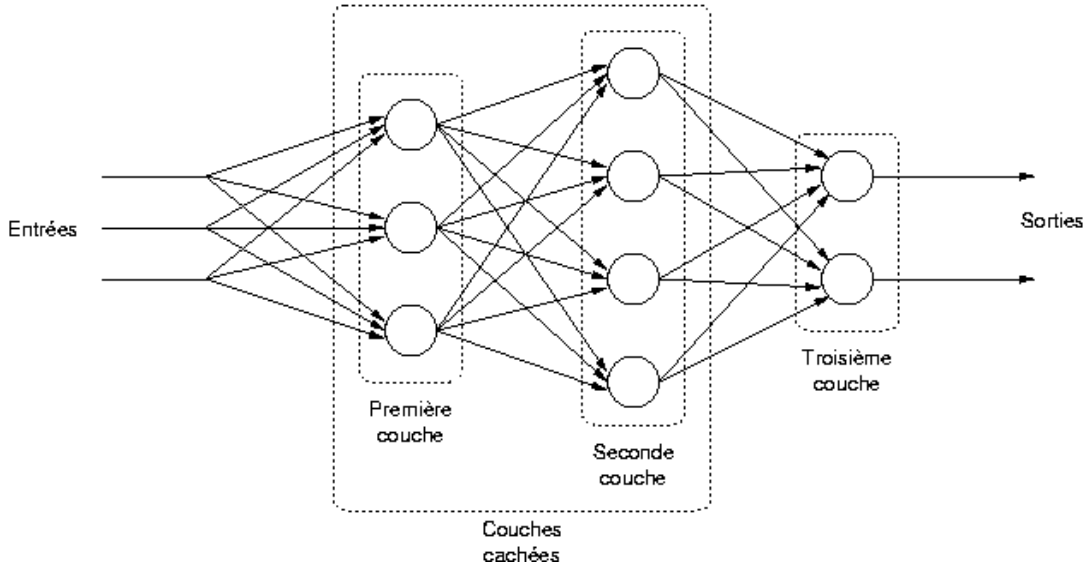


Figure 2-7: Schéma d'un réseau de neurones feedforward à trois couches, avec une couche d'entrée, une couche cachée et une couche de sortie.

2.6 Rappel de quelques travaux scientifiques sur le cancer du sein:

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

La Recherche sur l'algorithme de régression logistique des données de diagnostic du cancer du sein par apprentissage automatique a utilisé un algorithme de régression logistique pour classer l'ensemble de données de patientes atteintes d'un cancer du sein, l'ensemble de données a été obtenu du référentiel UCI Wisconsin. L'auteur a d'abord utilisé les 33caractéristiques de l'ensemble de données pour entraîner le modèle et a finalement obtenu une précision de **90 %**. Ensuite, l'auteur, à l'aide d'une technique de sélection de caractéristiques, a extrait deux caractéristiques principales des 33, à savoir la texture maximale et le périmètre maximal pour atteindre une précision de **96,5%,** ce qui est une amélioration par rapport au résultat obtenu à partir des 33 caractéristiques. En 2012,(**Yusuff, 2012)**

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

En utilisant Naïve Bayesian comme classificateur sur l'ensemble de données du Wisconsin de 10 caractéristiques **(Rashmi, 2016)**a essayé d'estimer le succès et l'erreur de l'algorithme de classification et de prédiction lorsque les données sont choisies au hasard.

Le modèle Naïve Bayes a montré un taux de réussite approximatif de 85 % à 95 % et un taux d'erreur de 10 à 15 % pour la classification et la prédiction.**(Amrane, 2018)**Dans leurs travaux de recherche, ont utilisé deux modèles d'apprentissage automatique : Naïve Bayesian et K Nearest voisin pour classer l'ensemble de données d'origine sur le cancer du sein du référentiel UCI Machine Learning.Leur objectif était de proposer lequel des deux est le plus efficace. En utilisant le même ensemble de données, ils y ont appliqué les différents algorithmes et en utilisant la validation croisée comme mesure de performance. Le résultat a montré que KNN avec **97,51%** pour la précision est légèrement meilleur que NB avec **96,19%.** Cependant, les auteurs ont suggéré qu'étant donné un ensemble de données plus important, le NB sera probablement plus performant parce que KNN sera affecté par sa complexité temporelle.

(Bazazeh & Shubair, 2016) Ont réalisé une étude comparative de trois algorithmes d'apprentissage automatique populaires pour la classification du cancer du sein : Support Vector Machine, Random Forest et Bayesian Network. Ils ont également utilisé l'ensemble de données original sur le cancer du sein du Wisconsin de l'UCI Machine Learning Repository. Les auteurs ont utilisé la technique de validation croisée K fold comme mesure de validation pour les classificateurs avec k = 10. Les paramètres utilisés pour leur comparaison étaient l'exactitude, la précision, le rappel et l'AUC ROC et après avoir effectué leur simulation sur l'ensemble de données avec les trois classificateurs, leur Le résultat montre que SVM a les performances les plus élevées en termes d'exactitude, de précision et de spécificité. Cependant, ils ont déclaré qu'en termes de classification correcte des tumeurs, le RF avait la probabilité la plus élevée.

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

2.7 Conclusion :

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

3 Chapitre 3 : Conception et Implémentation

3.1 Introduction :

Dans ce chapitre La première partie est une description de la base desdonnées, puis nous comprend le cadre mis en œuvre et les résultats avec le langage de programmation utilisé depuis la phase de prétraitement jusqu'à la phase d'apprentissage et de validation des modèles de prédiction. Des captures d'écran des résultats sont présentées pour appuyer notre cadre proposé.

3.2 Architecture du système :

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

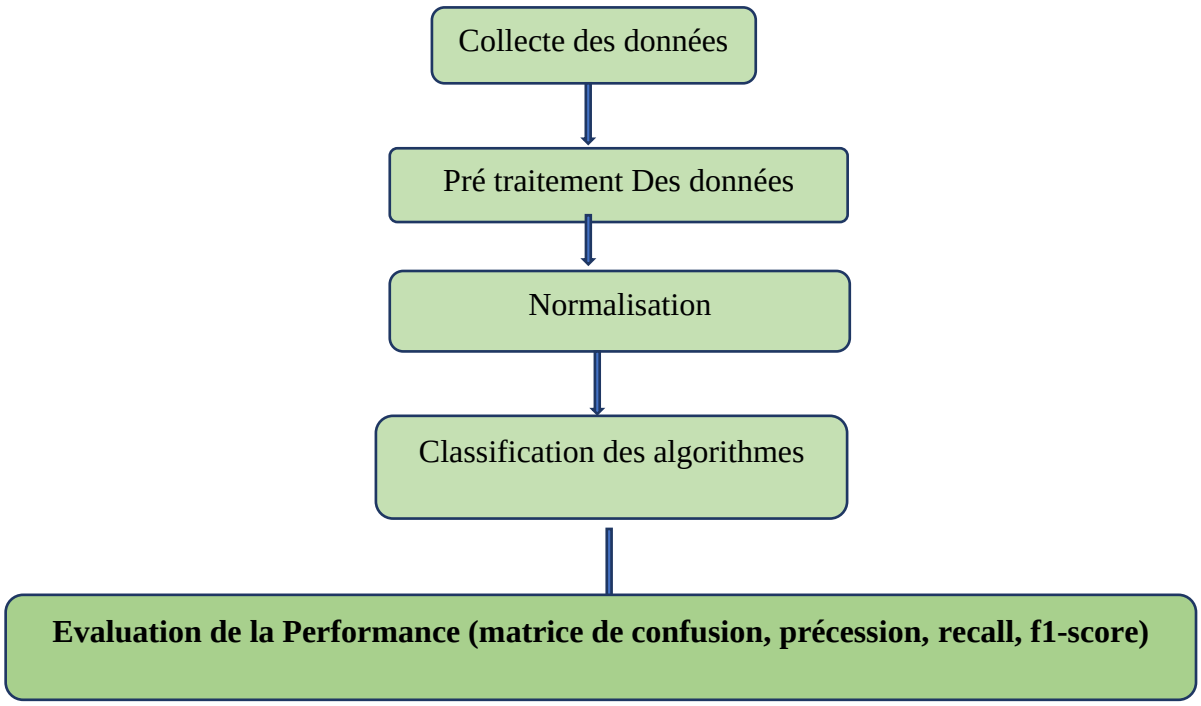


Figure 3-1 : Architecture de système pour l'apprentissage machine.

Le cadre proposé comprend les modules suivants : la collecte de données, l'étape de pré-traitement qui implique le traitement des données manquantes, la formation et le test des modèles d'apprentissage automatique et enfin, l'analyse et la comparaison des performances.
La **figure 3.1** illustre le cadre proposé, les données sont collectées à partir du référentiel d'apprentissage automatique en ligne de l'UCI. Les données collectées seront prétraitées, le prétraitement est effectué de manière à gérer les valeurs manquantes dans les données et une technique de mise à l'échelle des caractéristiques est utilisée pour normaliser les données. Les données sont divisées en un ensemble d'apprentissage (70 %) et un ensemble de test (30 %). La formation est utilisée pour former les quatre modèles de prédiction, tandis que l'ensemble de test est utilisé à des fins de validation.À l'aide de ces mesures de performance, qui sont la précision, le rappel et le score f1, les quatre modèles de prédiction sont évalués et comparés.

3.3 Les Outils Matériels et logiciels :

3.3.1 Partie logicielle :

Le modèle proposé a été implémenté à l'aide de google Colab avec un environnement de programmation python3, qui possède une bibliothèque d'apprentissage automatique (**SciKit Learn)**,L'un des principaux avantages de l'utilisation de Colab est qu'il contient la plupart des bibliothèques courantes nécessaires àl'apprentissage automatique tellesque (*TensorFlow*, *Keras*, *ScikitLearn*, *OpenCV*, *numpy*, *pandas*, etc). **SciKit Learn** prend en charge tous les algorithmes d'apprentissage automatique existants utilisés pour la classification, ainsi qu'un bon nombre de packages pour les techniques de prétraitement des données et les mesures de performance d'apprentissage automatique.
Ce langage présente des avantages majeurs par rapport aux autres en raison de sa flexibilité, étant donné la sortie après la convergence de l'étape d'apprentissage, la facilité de traçage des graphiques et des diagrammes.

3.3.2 Partie Matériel :

Google fournit également gratuitement le GPU (unité de traitement graphique) et le TPU (unité de traitement du tenseur). Ces accélérateurs matériels vous permettent d'exécuter des opérations d'apprentissage automatique lourdes sur de grands ensembles de données beaucoup

plus rapidement que n'importe quel environnement local. Colaboratory, souvent raccourci en "Colab", est un produit de Google Research. Colab permet à n'importe qui d'écrire et d'exécuter le code Python de son choix par le biais du navigateur. C'est un environnement particulièrement adapté au machine learning, à l'analyse de données et à l'éducation. En termes plus techniques, Colab est un service hébergé de notebooks Jupyter qui ne nécessite aucune configuration et permet d'accéder gratuitement à des ressources informatiques, dont des GPU. L'utilisation de Colab est gratuite.

3.4 Description de la base de données

Nom de la base de données : Wisconsin diagnostic breast cancer (WDBC)

Les informations pertinentes : Il existe un total de **569enregistrements** d'observation comprenant **357** cas bénins (B) et **212** cas malins (M). Le fichier de données est disponible sur le serveur de l'Université du Wisconsin à l'adresse http://ftp.cs.wisc.edu/math-prog/cpo-dataset/machine-learn/cancer/WDBC/WDBC.dat. Comme cette source manquait de noms de colonnes, les noms provenaient d'une version mise à jour qui ne pouvait être téléchargée que manuellement à l'adresse https://www.kaggle.com/uciml/breast-cancer-wisconsin-data. Il y a **33 colonnes** de données comprenant l'identification, le diagnostic et trois groupes de dix variables. Les valeurs statistiques moyennes pour les dix variables ci-dessous sont rapportées dans les colonnes 3-12.

Nombre de malade : 569

Nombre d'attributs : 33

Information des attributs :

- **ID nombre**
- **Diagnostique : (M=** malignant**, B=** benign**)**
- **Radius(rayon) : la distance entre le bord du cancer et le centre.**
- **Texture(texture) : écart type des valeurs d'échelle de gris.**
- **Perimeter (périmètre) : Est une mesure de taille du noyau cancéreux.**
- **Area(zone) : Est une mesure de la surface du cancer.**
- **Smoothness(douceur) : variation locale des longueurs de rayon.**
- **Compactness(compacité) : Est le rapport moyen entre le volume et la surface du cancer.**
- **Concavity (concavité) :Est le niveau moyen de concavité du contour du cancer.**

- **Concave points (points concaves) : Sont le nombre moyen de partie enfoncée du contour du cancer.**
- **Symmetry (symétrie) : Est le niveau de symétrie du cancer.**
- **Fractal dimension(dimension fractale) :approximation du littoral » - 1**

La moyenne, l'erreur standard et la valeur minimale de ces caractéristiques ont été calculés pour chaque donnée, ce qui a donné 30 caractéristiques. Par exemple, le champ 3 est le rayon moyen(Mean radius), le champ 13 est le rayon SE(Raduis se), le champ 23 est le pire rayon(Worst radius).

3.5 Prétraitement de la base des données :

Tous d'abord nous avons importé les librairies :

```
]  # importing libraries
   import pandas as pd
   import numpy as np
   import seaborn as sns
   import time
   from sklearn import svm
   import matplotlib.pyplot as plt
   import tensorflow as tf



   from sklearn.model_selection import train_test_split
   from sklearn.preprocessing import LabelEncoder
   from sklearn.preprocessing import StandardScaler
   from sklearn.linear_model import LogisticRegression
   from sklearn.metrics import accuracy_score
   from sklearn.metrics import classification_report
```

Figure 3-2 : : Représente le code python utilisé pour implémenter l'apprentissage du modèle de régression logistique.

Les données téléchargées à partir du référentiel d'apprentissage automatique UCI se trouvent sur ma machine locale dans ce répertoire "/content/drive/MyDrive/Memoire/rep 24 05 2021" avec le nom de fichier data.csv. La programmation Python a une bibliothèque connue sous le nom de Pandas, qui peut être utilisée pour ouvrir et lire des fichiers à valeurs séparées par des virgules (CSV) et la figure 3.2 montre le code python utilisé pour lire dans notre fichier de jeu de données

```
# charger la base de donéée
data = pd.read_csv("/content/drive/MyDrive/Memoire/rep 24 05 2021/data.csv")
print('Dimension des données : ', data.shape)
print( data.head())
```

```
Dimension des données :  (569, 33)
         id diagnosis  ...  fractal_dimension_worst  Unnamed: 32
0    842302         M  ...                  0.11890          NaN
1    842517         M  ...                  0.08902          NaN
2  84300903         M  ...                  0.08758          NaN
3  84348301         M  ...                  0.17300          NaN
4  84358402         M  ...                  0.07678          NaN

[5 rows x 33 columns]
```

Figure 3-3 : chargement de la base WDBC.

Quelque formule appliquer sur la base pour voir le contenue :

```
data
```

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavit
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	
...	...	...	...	...	...	...	...	...	
564	926424	M	21.56	22.39	142.00	1479.0	0.11100	0.11590	
565	926682	M	20.13	28.25	131.20	1261.0	0.09780	0.10340	
566	926954	M	16.60	28.08	108.30	858.1	0.08455	0.10230	
567	927241	M	20.60	29.33	140.10	1265.0	0.11780	0.27700	
568	92751	B	7.76	24.54	47.92	181.0	0.05263	0.04362	

569 rows × 33 columns

Activer Windows
Accédez aux paramètres pour activer Windows.

Figure 3-4 : : affichage de la base (WDBC).

On a deux types de données : float et int ,31 colonnes de type float et une colonne de type integer.

```
[80] data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 33 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   id                      569 non-null    int64
 1   diagnosis               569 non-null    object
 2   radius_mean             569 non-null    float64
 3   texture_mean            569 non-null    float64
 4   perimeter_mean          569 non-null    float64
 5   area_mean               569 non-null    float64
 6   smoothness_mean         569 non-null    float64
 7   compactness_mean        569 non-null    float64
 8   concavity_mean          569 non-null    float64
 9   concave points_mean     569 non-null    float64
 10  symmetry_mean           569 non-null    float64
 11  fractal_dimension_mean  569 non-null    float64
 12  radius_se               569 non-null    float64
 13  texture_se              569 non-null    float64
 14  perimeter_se            569 non-null    float64
 15  area_se                 569 non-null    float64
 16  smoothness_se           569 non-null    float64
 17  compactness_se          569 non-null    float64
 18  concavity_se            569 non-null    float64
 19  concave points_se       569 non-null    float64
 20  symmetry_se             569 non-null    float64
 21  fractal_dimension_se    569 non-null    float64
 22  radius_worst            569 non-null    float64
 23  texture_worst           569 non-null    float64
 24  perimeter_worst         569 non-null    float64
 25  area_worst              569 non-null    float64
 26  smoothness_worst        569 non-null    float64
 27  compactness_worst       569 non-null    float64
 28  concavity_worst         569 non-null    float64
 29  concave points_worst    569 non-null    float64
 30  symmetry_worst          569 non-null    float64
 31  fractal_dimension_worst 569 non-null    float64
 32  Unnamed: 32             0 non-null      float64
dtypes: float64(31), int64(1), object(1)
memory usage: 146.8+ KB
```

Activer Windows
Accédez aux paramètres pour activer Windows.

Figure 3-5 : Afficher plus informations sur la base (WDBC).

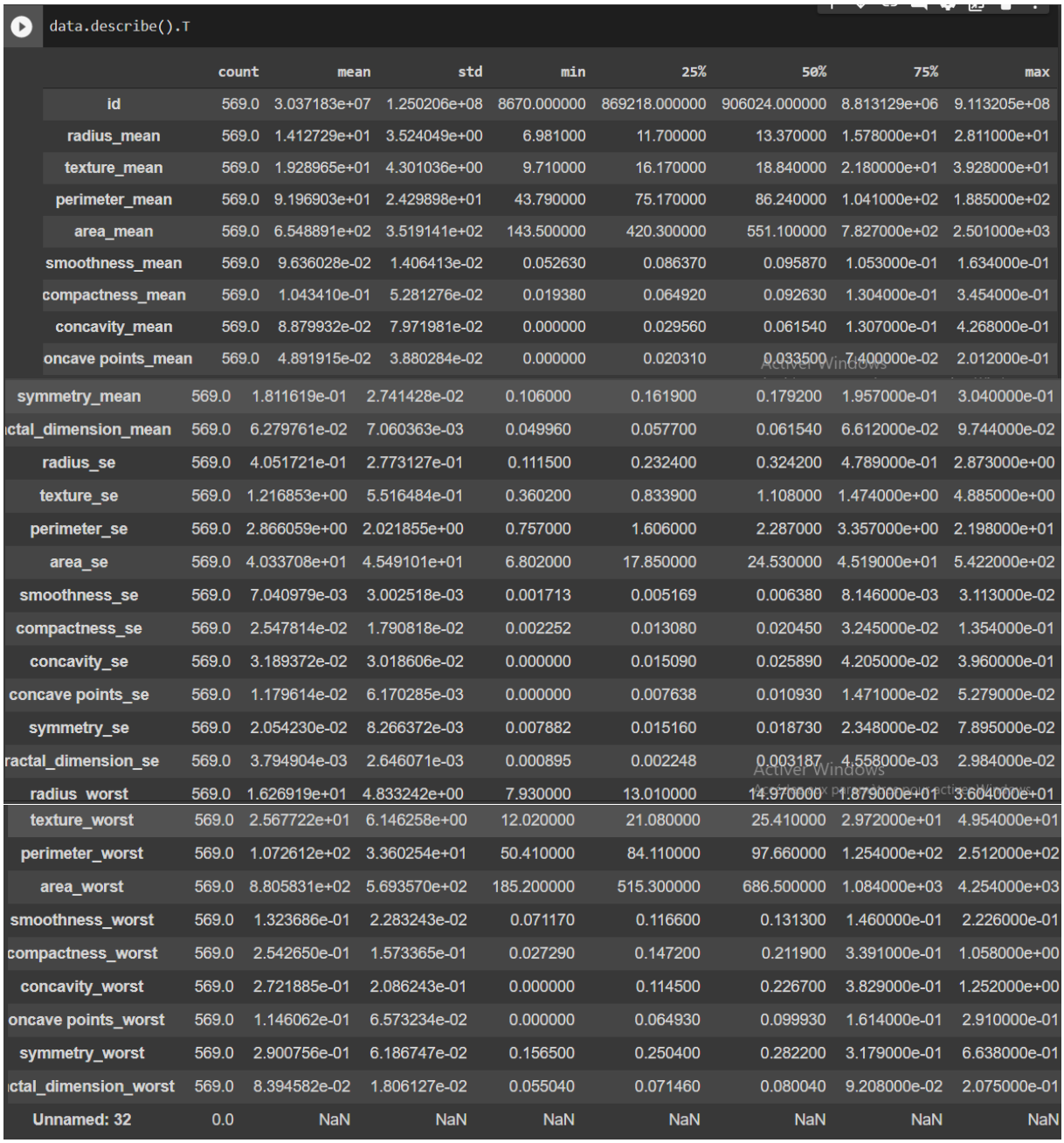
Pour La dimension de l'ensemble des données on a 569 lignes et 33 colonnes :

```
[19] data.shape

(569, 33)
```

Figure 3-6 : : Afficher la dimension du (WDBC).

On remarque que certaines patientes ont des données nulles et des données aberrantes pour quelques caractéristiques Les données nulles (les valeurs minimales de Conactivity_mean, Concave point_mean, Conactivity se, Concave point_se, Conactivity_Worst et concave point_Worst égales à 0).



```
data.describe().T
```

	count	mean	std	min	25%	50%	75%	max
id	569.0	3.037183e+07	1.250206e+08	8670.000000	869218.000000	906024.000000	8.813129e+06	9.113205e+08
radius_mean	569.0	1.412729e+01	3.524049e+00	6.981000	11.700000	13.370000	1.578000e+01	2.811000e+01
texture_mean	569.0	1.928965e+01	4.301036e+00	9.710000	16.170000	18.840000	2.180000e+01	3.928000e+01
perimeter_mean	569.0	9.196903e+01	2.429898e+01	43.790000	75.170000	86.240000	1.041000e+02	1.885000e+02
area_mean	569.0	6.548891e+02	3.519141e+02	143.500000	420.300000	551.100000	7.827000e+02	2.501000e+03
smoothness_mean	569.0	9.636028e-02	1.406413e-02	0.052630	0.086370	0.095870	1.053000e-01	1.634000e-01
compactness_mean	569.0	1.043410e-01	5.281276e-02	0.019380	0.064920	0.092630	1.304000e-01	3.454000e-01
concavity_mean	569.0	8.879932e-02	7.971981e-02	0.000000	0.029560	0.061540	1.307000e-01	4.268000e-01
oncave points_mean	569.0	4.891915e-02	3.880284e-02	0.000000	0.020310	0.033500	7.400000e-02	2.012000e-01
symmetry_mean	569.0	1.811619e-01	2.741428e-02	0.106000	0.161900	0.179200	1.957000e-01	3.040000e-01
ctal_dimension_mean	569.0	6.279761e-02	7.060363e-03	0.049960	0.057700	0.061540	6.612000e-02	9.744000e-02
radius_se	569.0	4.051721e-01	2.773127e-01	0.111500	0.232400	0.324200	4.789000e-01	2.873000e+00
texture_se	569.0	1.216853e+00	5.516484e-01	0.360200	0.833900	1.108000	1.474000e+00	4.885000e+00
perimeter_se	569.0	2.866059e+00	2.021855e+00	0.757000	1.606000	2.287000	3.357000e+00	2.198000e+01
area_se	569.0	4.033708e+01	4.549101e+01	6.802000	17.850000	24.530000	4.519000e+01	5.422000e+02
smoothness_se	569.0	7.040979e-03	3.002518e-03	0.001713	0.005169	0.006380	8.146000e-03	3.113000e-02
compactness_se	569.0	2.547814e-02	1.790818e-02	0.002252	0.013080	0.020450	3.245000e-02	1.354000e-01
concavity_se	569.0	3.189372e-02	3.018606e-02	0.000000	0.015090	0.025890	4.205000e-02	3.960000e-01
concave points_se	569.0	1.179614e-02	6.170285e-03	0.000000	0.007638	0.010930	1.471000e-02	5.279000e-02
symmetry_se	569.0	2.054230e-02	8.266372e-03	0.007882	0.015160	0.018730	2.348000e-02	7.895000e-02
ractal_dimension_se	569.0	3.794904e-03	2.646071e-03	0.000895	0.002248	0.003187	4.558000e-03	2.984000e-02
radius_worst	569.0	1.626919e+01	4.833242e+00	7.930000	13.010000	14.970000	1.879000e+01	3.604000e+01
texture_worst	569.0	2.567722e+01	6.146258e+00	12.020000	21.080000	25.410000	2.972000e+01	4.954000e+01
perimeter_worst	569.0	1.072612e+02	3.360254e+01	50.410000	84.110000	97.660000	1.254000e+02	2.512000e+02
area_worst	569.0	8.805831e+02	5.693570e+02	185.200000	515.300000	686.500000	1.084000e+03	4.254000e+03
smoothness_worst	569.0	1.323686e-01	2.283243e-02	0.071170	0.116600	0.131300	1.460000e-01	2.226000e-01
compactness_worst	569.0	2.542650e-01	1.573365e-01	0.027290	0.147200	0.211900	3.391000e-01	1.058000e+00
concavity_worst	569.0	2.721885e-01	2.086243e-01	0.000000	0.114500	0.226700	3.829000e-01	1.252000e+00
oncave points_worst	569.0	1.146062e-01	6.573234e-02	0.000000	0.064930	0.099930	1.614000e-01	2.910000e-01
symmetry_worst	569.0	2.900756e-01	6.186747e-02	0.156500	0.250400	0.282200	3.179000e-01	6.638000e-01
ctal_dimension_worst	569.0	8.394582e-02	1.806127e-02	0.055040	0.071460	0.080040	9.208000e-02	2.075000e-01
Unnamed: 32	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figure 3-7 : Afficher les données Nulles.

Cet Histogramme représente le nombre de personnes atteintes d'une tumeur sont 212, et le nombre de personnes qui ne sont pas cancéreux sont 357.

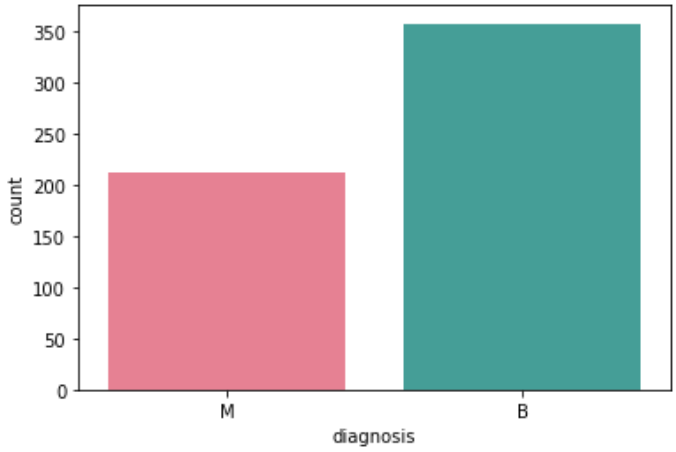


Figure 3-8 : Afficher histogramme (bénigne ou malin).

Une matrice de corrélation est utilisée pour examiner la dépendance entre plusieurs variables en même temps. Le résultat est une matrice contenant les coefficients de corrélation entre chaque variable et les autres. Il existe différentes méthodes de test de corrélation à utiliser.

Une corrélation entre les variables indique que lorsqu'une variable change de valeur, l'autre variable a tendance à changer dans une direction spécifique. Les nuages de points et la carte thermique sont de bonnes techniques de visualisation pour visualiser la corrélation entre les variables continues comme la **figure 3.9.**

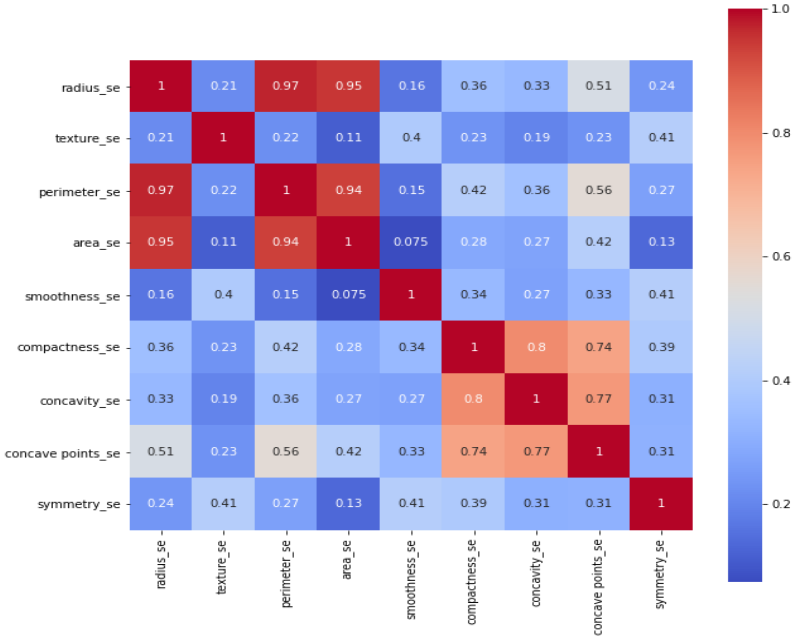


Figure 3-9 : Matrice de corrélation de SE _Colonnes.

Faire la codification de la classe qui contient un (M,B) par 1 et 0 respectivement :

```
#codification  M par 1 et B par 0
encoder = LabelEncoder ()
y = encoder.fit_transform(y)
print('Aprés le codage :', y )
```

```
Aprés le codage : [1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 0 1 1 1 1 1 1 1 0 1 0 0 0 0 0 1 1 0 1 1 0 0 0 0 1 0 1 1 0 0 0 0 1 0 1 1
 0 1 0 1 1 0 0 0 1 1 0 1 1 1 0 0 0 1 0 0 1 1 0 0 0 1 1 0 0 0 0 1 0 0 1 0 0
 0 0 0 0 0 0 1 1 1 0 1 1 0 0 0 1 1 0 1 0 1 1 0 1 1 0 0 1 0 0 1 0 0 0 0 1 0
 0 0 0 0 0 0 0 0 1 0 0 0 0 1 1 0 1 0 0 1 1 0 0 1 1 0 0 0 0 1 0 0 1 1 1 0 1
 0 1 0 0 0 1 0 0 1 1 0 1 1 1 1 0 1 1 1 0 1 0 1 0 0 1 0 1 1 1 1 0 0 1 1 0 0
 0 1 0 0 0 0 0 1 1 0 0 1 0 0 1 1 0 1 0 0 0 0 1 0 0 0 0 0 1 0 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 0 0 0 0 0 0 1 0 1 0 0 1 0 0 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0
 0 1 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 1 0 0 0 0 1 1 1 0 0
 0 0 1 0 1 0 1 0 0 0 1 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1
 1 0 1 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 1 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0
 0 1 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 0 1 0 0 0 0 0 1 0 0
 1 0 1 0 0 1 0 1 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0
 0 0 0 0 0 0 1 0 1 0 0 1 0 0 0 0 0 0 1 1 0 1 0 1 0 0 0 0 0 1 0 0 1 0 1 0 1 1
 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 1 1 1 1 1 1 0]
```

Figure 3-10 : la codification de la classe M par 1 et B par 0.

Le code python utilisé pour implémenter la mise à l'échelle des fonctionnalités de l'ensemble de données à l'aide d'une méthode StandardScalar dans la classe de prétraitement de la bibliothèque SciKit Learn.

```
[ ]  # Scaling nos données d'apprentissage
     # chaque valeur des attributs aura une moyenne 0 et standard deviation égale à 1
     sc = StandardScaler()

[ ]  sc.fit(x_train)

     StandardScaler(copy=True, with_mean=True, with_std=True)

[ ]  x_test=sc.transform(x_test)

[ ]  x_train=sc.transform(x_train)
```

Figure 3-11 : Mise à l'échelle des fonctionnalités de WBCD.

Pour la classification, les données prétraitées sont introduites dans le modèle de régression logistique. Les données sont divisées en deux parties ; l'ensemble d'apprentissage (70 % des données) et l'ensemble de test (30 % des données). Pour entraîner le modèle, nous avons utilisé l'ensemble d'apprentissage et pour évaluer les performances du modèle, nous avons utilisé l'ensemble de test. **La figure 3.12**montre le code de programmation Python utilisé pour implémenter cela.

```
[ ]  x_train, x_test, y_train, y_test = train_test_split(x,y,test_size=0.3,random_state = 40)

[ ]  # partage des données en données des tests et d'apprentissage
     # Données d'apptrentissage : 70 % (X_train et Y_train)
     # Donnes de tests : 30 % (X_test et Y_test)
     x = data.iloc[:,2:].values

[ ]  #y = data.iloc[:,1:1].values
     x = data.iloc[:,1:10]
     y = data.iloc[:,10:11]
```

Figure 3-12 : WDBC divisé en deux partie (apprentissage, teste).

3.6 Classification avec le modèle d'Apprentissage :

3.6.1 Matrice de confusion :

Une matrice de confusion inclut des informations sur les classifications réelles et prévues effectuées par un modèle de classification. Les performances d'un tel modèle sont généralement évaluées à l'aide des données de la matrice.

Le **tableau 1** montre la matrice de confusion pour un classificateur à deux classes (Goyal & Mehta, 2012). Il classe chaque instance dans l'une des deux classes.

Les classes sont vraies et fausses ; cela donne lieu à quatre classifications possibles pour chaque instance, comme indiqué ci-dessous.

Vrai-Positive (VP) signifie un modèle positif considéré comme positif.

Faux-positif (FP) signifie un modèle négatif considéré comme positif.

Faux-négatif (FN) signifie positif Modèle considéré comme négatif.

Vrai-négatif (VN) signifie un motif négatif considéré comme négatif.

D'après le tableau ci-dessous, la classification qui se trouve le long de la diagonale principale qui est TP et TN sont les classifications/prédictions correctes.Alors que les champs restants, FN et FP signifient une erreur de modèle.

		Classe réel	
		Positif (0)	Négatif (1)
Classe prévues	Positive (0)	Vrai positif (VP)	Faux négative (FN)
	Négative (1)	Faux positive (FP)	Vrai négative (VN)

Tableau 1 : Matrice de confusion pour le classificateur à deux classes.

3.6.2 Les métriques :

L'ensemble des métriques sont utilisées pour évaluer les méthodes d'apprentissage automatique. À partir de la matrice de confusion, de nombreuses mesures de performance du modèle peuvent être dérivées, parmi lesquelles la précision est la plus populaire, qui est définie comme :

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

$$\boldsymbol{Accuracy} = \frac{\mathrm{VP+VN}}{\mathrm{VP+FP+FN+VN}}$$

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

$$\textbf{Précision} = \frac{\mathbf{VP}}{\boldsymbol{VP+FP}} \quad \text{pour cas bénigne}$$

$$\textbf{Précision} = \frac{\mathbf{VN}}{\boldsymbol{VN+FN}} \quad \text{pour cas malin}$$

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

$$\textbf{Recall} = \frac{\mathbf{VP}}{\boldsymbol{VP+FN}} \quad \text{pour cas bénigne}$$

$$\textbf{Recall} = \frac{\textbf{VN}}{\boldsymbol{VN+FP}} \quad \text{pour cas malin}$$

F1-Score : Cela traduit l'équilibre entre la précision et le rappel ; c'est la moyenne harmonique de Precision et Recall

$$\textbf{F1-Score} = \frac{\textbf{2}}{\left(\frac{\textbf{1}}{\boldsymbol{precesion}}\right)+\left(\frac{\textbf{1}}{\boldsymbol{Recall}}\right)}$$

3.7 Implémentation des modèles :

Cette figure Représente le code python utilisé pour implémenter l'apprentissage du modèle de régression logistique.

```
[62] # instance (our model) of LogisticRegression is created
     lr=LogisticRegression()
     model1=lr.fit(x_train,y_train)
     pred1=model1.predict(x_test)
```

Figure 3-13 : modèle de LR.

Cette figure Représente le code python utilisé pour implémenter l'apprentissage du modèle de Naïve Bayes.

```
[38] #GaussianNB class imported from sklearn.naive_bayes
     from sklearn.naive_bayes import GaussianNB

[39] #instance(our model) of naiv_bayes is created
     gnb =GaussianNB()
     gnb.fit(x_train,y_train)
     pred3=gnb.predict(x_test)
```

Figure 3-14 : modèle de NB.

Cette figure Représente le code python utilisé pour implémenter l'apprentissage du modèle de Support Vecteur Machine.

```
#LinearSVC class imported from sklearn.svm
from sklearn.svm import LinearSVC
```

```
[33] svc = svm.SVC()
     svc.fit(x_train,y_train)
```

Figure 3-15 : modèle de SVM.

Cette figure Représente le code python utilisé pour implémenter l'apprentissage du modèle de K-Nearest -Neighbors.

```
[51] from sklearn.neighbors import  KNeighborsClassifier
```

```
[52] kNN = KNeighborsClassifier ()
     kNN.fit(x_train,y_train)
     pred5=kNN.predict(x_test)
```

Figure 3-16 : modèle de KNN.

Cette figure Représente le code python utilisé pour implémenter l'apprentissage du modèle de Multiplayer-Perceptron.

```
[44] #MLPClassifier class imported from sklearn.neural_network
     from sklearn.neural_network import  MLPClassifier
```

```
[45] #instance(our model)of MLPClassifier is created
     mlp_clf=MLPClassifier(solver='adam',alpha=1e5,max_iter=10000,hidden_layer_sizes=(5,2),random_state=1)
```

```
[46] #Using the fit method the model is trained
     mlp_clf.fit(x_train,y_train)
     pred4=mlp_clf.predict(x_test)
```

Figure 3-17: : modèle de MLP

Après la partie d'apprentissage et de training l'étape suivante consiste à tester l'intelligence du modèle, à cette fin, les données de test sont utilisées.L'ensemble de test comporte **569** points de données avec 33 caractéristiques indépendantes et une étiquette cible. Pour tester le modèle entraîné, l'ensemble de test à l'exclusion de l'étiquette cible est transmis au modèle pour que le modèle fasse, Certaines prédictions. Les prédictions (résultat prévu) du modèle seront

utilisées pour correspondre aux résultats réels de l'ensemble de test. Ces figures montrent le code python utilisé pour tester les modèles et les résultats prédits sont affichés.

Prédiction de LR :

```
[68] y_pred1 = lr.predict(x_test)

[69] y_pred1

array([0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1,
       0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0,
       1, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1,
       1, 1, 0, 0, 0, 1, 1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 1,
       0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0,
       0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1,
       0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1])
```

Figure 3-18: Le modèle de LR est testé

Prédiction de SVM :

```
[71] y_pred2 =svc.predict(x_test)
     y_pred2

     array([0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1,
            0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0,
            1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1,
            1, 1, 0, 0, 0, 1, 1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 1,
            0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0,
            0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1,
            0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1])
```

Figure 3-19 : Le modèle de SVM est testé.

Prédiction de NB :

```
[72] y_pred3 = gnb.predict(x_test)
     y_pred3

     array([0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1,
            0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0,
            1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 1,
            1, 1, 0, 0, 0, 1, 1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 1,
            0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0,
            0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1,
            0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1])
```

Figure 3-20 : Le modèle de NB est testé.

Prédiction de KNN :

```
[77] y_pred5 = kNN.predict(x_test)
     y_pred5

     array([0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1,
            0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0,
            1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1,
            1, 1, 0, 0, 0, 1, 1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 1,
            0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0,
            0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1,
            0, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1])
```

Figure 3-21 : Le modèle de KNN est testé.

Prédiction de MLP :

```
[75] y_pred4 = mlp_clf.predict(x_test)
     y_pred4

     array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0])
```

Figure 3-22 : Le modèle de MLP est testé.

3.8 Analyse de performance :

Quatre mesures de performance, à savoir **la matrice de confusion**, **la précision**, **le rappel (recall)** et **le score f1** : sont utilisées pour évaluer les performances des modèles entraînés. Ensuite, leurs performances sont discutées, analysées et des hypothèses sont faites. La figure 4.12 montre la matrice de confusion, le score de précision, le score de précision, le score de rappel et le score f1 pour les modèles (LR, SVM, NB, MLP, KNN) lorsque la mise à l'échelle des caractéristiques n'est pas appliquée sur le WBCD :

```
[37] from sklearn.metrics import confusion_matrix
     cm = confusion_matrix (y_test,pred1)
     cm

     array([[113,   2],
            [  2,  54]])
```

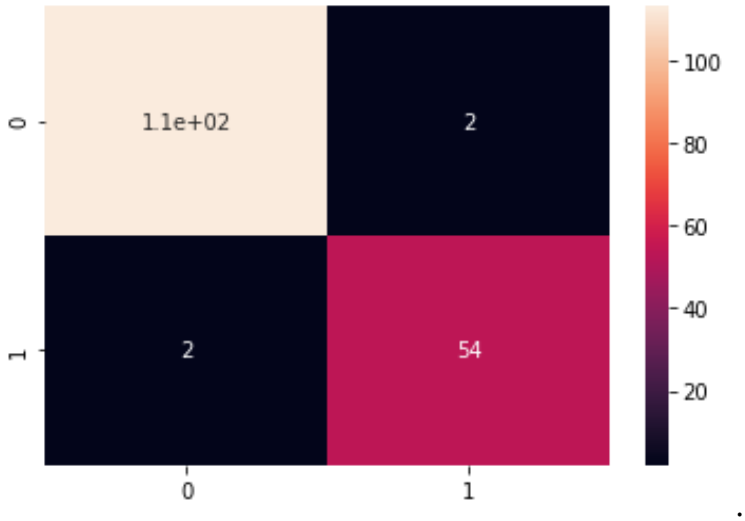


Figure 3-23 : La matrice de confusion de modèle LR.

Quatre mesures de performance, à savoir la matrice de confusion, la précision, le rappel et le score f1, sont utilisées pour évaluer les performances des modèles entraînés. Ensuite, leurs performances sont discutées, analysées et des hypothèses sont faites. La figure 3.23montre la matrice de confusion, le score de précision, le score de précision, le score de rappel et le score f1 pour le modèle LR lorsque la mise à l'échelle des caractéristiques n'est pas appliquée sur le WBCD.

```
TP=cm[0][0]
TN=cm[1][1]
FN=cm[1][0]
FP=cm[0][1]
print('Testing Accuracy:',(TP+TN)/(TP+TN+FP+FN))

Testing Accuracy: 0.9766081871345029

[73] accuracy_score(y_test,pred1)

     0.9766081871345029

print(classification_report(y_test,pred1))

                precision    recall  f1-score   support

             0       0.98      0.98      0.98       115
             1       0.96      0.96      0.96        56

      accuracy                           0.98       171
     macro avg       0.97      0.97      0.97       171
  weighted avg       0.98      0.98      0.98       171
```

Figure 3-24 : Les Métriques de performance pour le modèle LR.

Quatre mesures de performance, à savoir la matrice de confusion, la précision, le rappel et le score f1, sont utilisées pour évaluer les performances des modèles entraînés. Ensuite, leurs performances sont discutées, analysées et des hypothèses sont faites. La figure 3.25montre la matrice de confusion, le score de précision, le score de précision, le score de rappel et le score f1 pour le modèle SVM lorsque la mise à l'échelle des caractéristiques n'est pas appliquée sur le WBCD.

```
[45] pred2=svc.predict(x_test)
     cm1 = confusion_matrix (y_test,pred2)
     cm1

     array([[114,   1],
            [  1,  55]])

     sns.heatmap(cm1,annot=True)
     plt.savefig('h.png')
```

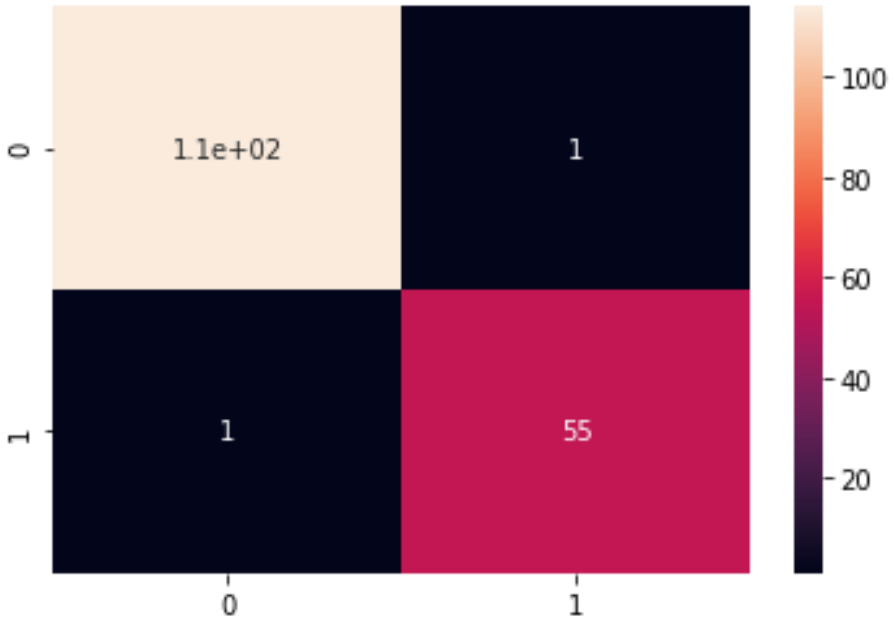


Figure 3-25 : La matrice de confusion de modèle SVM.

```
[47] print(accuracy_score(y_test,pred2))

     0.9883040935672515

[48] print(classification_report(y_test,pred2))

                  precision    recall  f1-score   support

               0       0.99      0.99      0.99       115
               1       0.98      0.98      0.98        56

        accuracy                           0.99       171
       macro avg       0.99      0.99      0.99       171
    weighted avg       0.99      0.99      0.99       171
```

Figure 3-26 : Métriques de performance pour le modèle SVM.

Quatre mesures de performance, à savoir la matrice de confusion, la précision, le rappel et le score f1, sont utilisées pour évaluer les performances des modèles entraînés. Ensuite, leurs performances sont discutées, analysées et des hypothèses sont faites. La **figure 3.27** montre la matrice de confusion, le score de précision, le score de précision, le score de rappel et le score f1 pour le modèle NB lorsque la mise à l'échelle des caractéristiques n'est pas appliquée sur le WBCD.

```
[52] cm2 = confusion_matrix (y_test,pred3)
     cm2

     array([[112,   3],
            [  3,  53]])

[53] sns.heatmap(cm2,annot=True)
     plt.savefig('h.png')
```

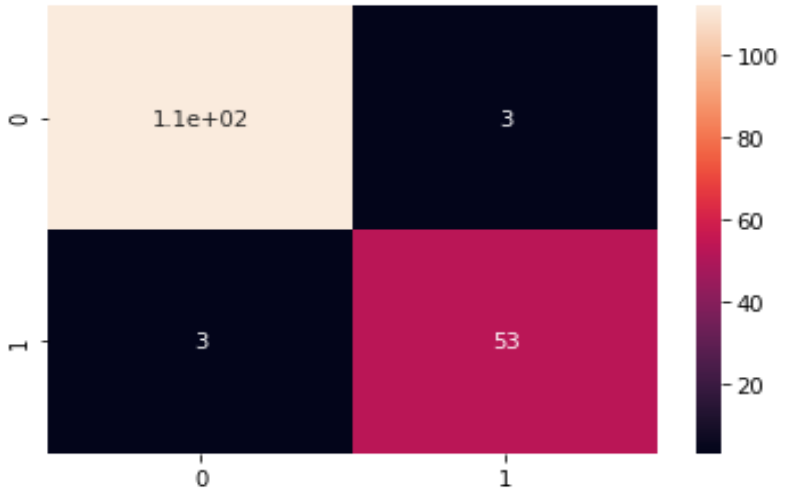


Figure 3-27 : La matrice de confusion de modèle NB.

```
[42] #the prediction accuracy of the trained model is tested
     print(accuracy_score(y_test,pred3))

     0.9590643274853801

[43] print(classification_report(y_test,pred3))

                   precision    recall  f1-score   support

                0       0.97      0.97      0.97       115
                1       0.93      0.95      0.94        56

         accuracy                           0.96       171
        macro avg       0.95      0.96      0.95       171
     weighted avg       0.96      0.96      0.96       171
```

Figure 3-28 : Métriques de performance pour le modèle NB.

Quatre mesures de performance, à savoir la matrice de confusion, la précision, le rappel et le score f1, sont utilisées pour évaluer les performances des modèles entraînés. Ensuite, leurs performances sont discutées, analysées et des hypothèses sont faites. La **figure 3.29**montre la matrice de confusion, le score de précision, le score de précision, le score de rappel et le score f1 pour le modèle MLP lorsque la mise à l'échelle des caractéristiques n'est pas appliquée sur le WBCD.

```
[ ] cm3 = confusion_matrix (y_test,pred4)
    cm3

    array([[115,   0],
           [ 56,   0]])
```

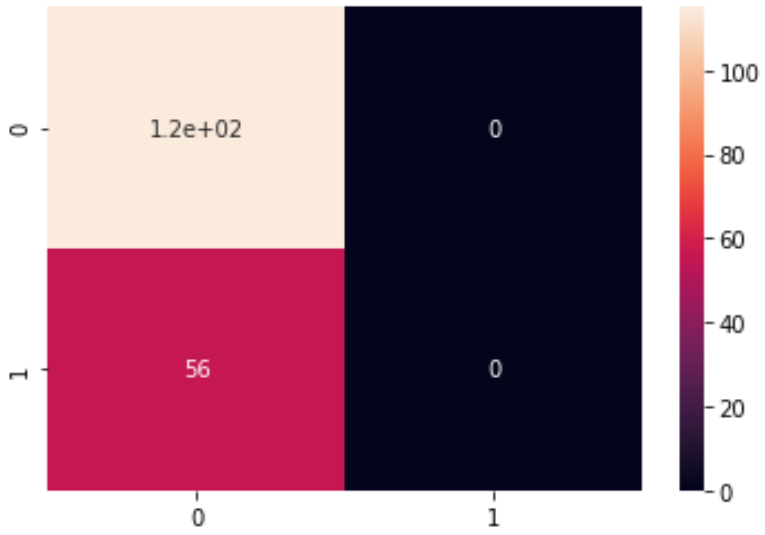


Figure 3-29 : La matrice de confusion de MLP.

```
[ ]  print(accuracy_score(y_test,pred4))

     0.672514619883041

     print(classification_report(y_test,pred4))

                   precision    recall  f1-score   support

                0       0.67      1.00      0.80       115
                1       0.00      0.00      0.00        56

         accuracy                           0.67       171
        macro avg       0.34      0.50      0.40       171
     weighted avg       0.45      0.67      0.54       171
```

Figure 3-30 : : Métriques de performance pour le modèle MLP.

Quatre mesures de performance, à savoir la matrice de confusion, la précision, le rappel et le score f1, sont utilisées pour évaluer les performances des modèles entraînés. Ensuite, leurs performances sont discutées, analysées et des hypothèses sont faites. La **figure 3.31**montre la matrice de confusion, le score de précision, le score de précision, le score de rappel et le score f1 pour le modèle KNN lorsque la mise à l'échelle des caractéristiques n'est pas appliquée sur le WBCD.

```
[ ]  cm4 = confusion_matrix (y_test,pred5)
     cm4

     array([[113,   2],
            [  2,  54]])
```

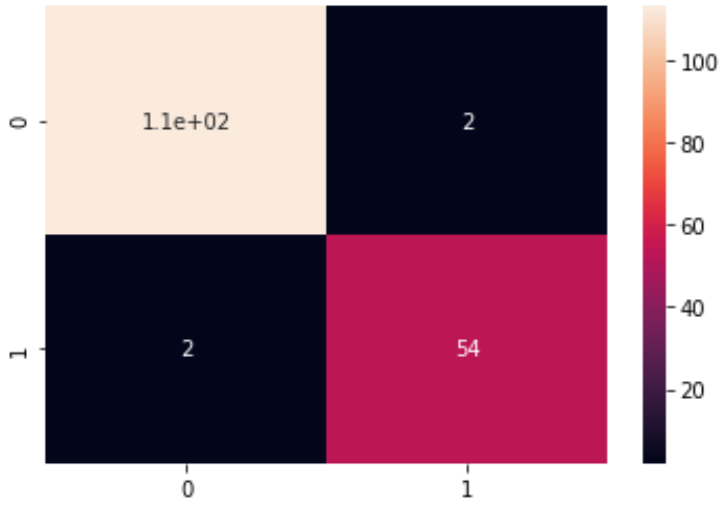


Figure 3-31 : La matrice de confusion de KNN.

```
[ ] print(accuracy_score(y_test,pred5))

    0.9766081871345029

[ ] print(classification_report(y_test,pred5))

                  precision    recall  f1-score   support

               0       0.98      0.98      0.98       115
               1       0.96      0.96      0.96        56

        accuracy                           0.98       171
       macro avg       0.97      0.97      0.97       171
    weighted avg       0.98      0.98      0.98       171
```

Figure 3-32 : Métriques de performance pour le modèle KNN.

Performance			
Model	Precision	Recall	F1-Score
LR	97%	97%	97%
SVM	99%	99%	99%
NB	95%	96%	97%
MLP	34%	50%	40%
KNN	97%	97%	97%

Tableau 2: Métrique de performance.

Model	LR	SVM	NB	MLP	KNN
Accuracy	98%	99%	96%	67%	98%

Tableau 3 : Résultat Accuracy.

Le classificateur Support Vector Machine (99%) estplus performant en termes de précision et de spécificité. En outre, il est également à noter que Naïve bayes (96%) a donné la sensibilité la plus petite. Mais par rapport aux autres métriques de performances,

L'étude d'amélioration des performances basée sur l'analyse et la modification de WDBC. Comme le classificateur NB fonctionnait mieux parmi nos classificateurs

proposés, nous avons essayé d'optimiser davantage le résultat. Nous avons essayé de trouver l'efficacité de chaque fonctionnalité et leurs effets sur les performances

3.6 Conclusion :

Dans ce dernier chapitre nous avons fait la description de la base de données WDBC, après nous allons appliquer cinq méthodes de performances. Ensuite nous comparons ces algorithmesentre eux. Pour avoir un résultat Les modèles en affichant le rapport et accuracy de chaque modèle, on peut choisir quel modèle performe le mieux. Pour le choisir, il faut se baser sur l'accuracy la plus adéquate.

Conclusion générale

Dans ce mémoire, nous avons présenté des modèles d'apprentissage automatique supervisé pour la classification de cancer du sein sur l'ensemble des données de la base WDBC.

On ne saurait trop insister sur la nécessité d'un prédicteur précis pour la prédiction du cancer du sein. Le cancer du sein a le deuxième taux de mortalité le plus élevé, où le cancer du poumon est le premier et ce cancer touche principalement les femmes. Pour sa détection et sa classification, les médecins ont utilisé la mammographie pour établir un diagnostic sur leurs patientes. Cependant, la précision de la mammographie est moins impressionnante, de sorte que le besoin d'un meilleur facilitateur de prédiction est de plus en plus pressant. De nombreux chercheurs ont utilisé les techniques d'apprentissage automatique et d'intelligence artificielle pour la prédiction et la classification du cancer du sein.

Pour la tâche de classification, les données sont divisées en deux ensembles, qui sont l'ensemble d'apprentissage (70 % des données) et l'ensemble de test (30 % des données). Le classifier SVM a donné le meilleur résultat (99%).

Par conséquent le modèle SVMpeut être utilisé pour la prédiction du cancer du sein ce qui aide grandement les médecins à établir un diagnostic approprié.
Ce travail peut être amélioré comme suit :
Proposer d'autre modèle pour la classification du cancer du sein.

Bibliographie

A. Cornuéjols, L. Miclet, Y.Kodratoff. 2002. *apprentrissage artificiel,concepts et algorithmes.* 2002.

Agrawal, Gunopulos et Leymann, sd ,. 2004. s.l. : Tao, Faloutsos, Papadias et Liu, 2004.

Agrawal, Gunopulos et Leymann, sd et Tao, Faloutsos, Papadias et Liu,. 2004. 2004.

Amrane, Oukid, Gagaoua et Ensari,. 2018. 2018.

Bahaj, Khourdifi &. 2019. 2019.

Bazazeh & Shubair. 2016. 2016.

Bellaterra, Edicions. 2013. *Cuerpos de cine : masculinidades carnales en el cine y la cultura popular contemporáneos.* 2013.

BENZAKI, Younes. 2016-2017. *Naive Bayes Classifier pour la Classification en Machine .* 2016-2017.

Bhaya, Dr. Wesam S. 2017. *Review of Data Preprocessing.* 2017.

Brahim, Boughaba Mohammed et Boukhris. 2016-2017. *L`apprentissage profond (Deep Learning).* ouergla : s.n., 2016-2017.

Chaumet, Hélène. 2019. *Cancer du sein et de l'ovaire.* 2019.

Dana Bazazeh, Raed Shubair. 2016. *Comparative study of machine learning algorithms for breast cancer detection and diagnosis.* 2016.

Desponds, Elodie. 2014. *L'accompagnement infirmier auprès des femmes de moins de cinquante ans, subissant un impact psychosocial, dans les jours suivant l'annonce du diagnostic du cancer du sein.* 2014. oai:doc.rero.ch:20150204103749-IV.

Godé, Hadrien. 2020. *Étude de modèles de substitution, application à la.* s.l. : Centre Européen de Recherche et de Formation Avancée du Calcul, 2020.

Gupta & Gupta. 2018. 2018.

Issy-les-Moulineaux, Elsevier Masson. 2018. *tumeurs du sein.* france : s.n., 2018.

Jason. 2013. 2013.

Karpagavalli, S, Jamuna, K S et Vijaya, M S. 2009. *International Journal of Recent Trends in Engineering; Oulu.* 2009.

Learned-Miller. 2014. *Introduction to Supervised Learning.* s.l. : Department of Computer Science Amherst, MA 01003, 2014.

Md. Toukir Ahmed, Md. Niaz Imtiaz and Animesh Karmakar. 2020. *Analysis of Wisconsin Breast Cancer original dataset using data.* 2020.

Mohamed Bouguessa, Lotfi Ben Romdhane. 2015. *ACM Transactions on Intelligent Systems and Technology.* 2015. 30.

Rashmi, Lekha et Bawane. 2016. 2016.

Sung, Charan GudlaAndrew H. 2019. *Evaluating Machine Learning Models on the Ethereum Blockchain for Android Malware Detection.* 2019.

Vandenbossche, Dr Gautier. 2016. *Mieux comprendre le cancer du sein.* 2016.

Wikipedia., The original uploader was Vlad2i at French. 2005. *Sigmoide.* 2005.

wiliam. 2020. *Alimentation et nutrition : quels effets sur notre santé ?* 2020.

Winslow., Terese. 2018. *Medical And Scientific Illustration.* 2018.

Yusuff, Mohamad, Ngah et Yahaya. 2012. 2012.