République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université de Mohamed El Bachir El Ibrahimi de Bordj Bou Arréridj
Faculté des Mathématiques et d'Informatique
Département d'informatique

MEMOIRE

Présenté en vue de l'obtention du diplôme

Master en informatique

Spécialité : Technologies de l'Information et de la Communication

THEME

La prédiction des Maladies Basée sur les Symptômes à l'Aide de l'apprentissage Automatique

Présenté par :

BENSACI SAHRA

KHRAMSSIA NOUR EL HOUDA

Soutenu publiquement le : 22/06/2023

Devant le jury composé de :

Président :Mr : Beghoura Mohamed Amine.

Examinateur : Mme : Zitouni Sihem.

Encadreur : Mr : Nouioua Mourad.

2022/2023

Dédicace

Avec l'expression de ma reconnaissance,

Je dédie ce modeste travail :

Mes chers parents, qui mont dotée d'une éducation digne, leur amour a fait de moi ce que je suis aujourd'hui.

Mes frères Amine, Hamza, Taki el dine et mes sœurs Imane, Chaima.

Mes chers amis, pour leur appui et leur encouragement. Un merci particulier à mon collègue Mehiris Hicham pour son soutien et sa présence.

Toute ma famille, pour leur soutien tout au long de mon parcours universitaire. Que ce travail soit l'accomplissement de vos vœux tant allégués et le fruit de votre soutien infaillible. Merci d'être toujours là pour moi.

Un grand merci à ma collègue Sahra, avec qui j'ai partagé les meilleurs moments durant la réalisation de ce travail.

Nour el Houda

Dédicace

Je dédie ce travail A mon père et ma mère en témoignage de leur affectation, leurs sacrifices et de leurs précieux conseils qui m'ont conduit à la réussite dans mes études.

A mes frère Islam et Hocine, ma sœur Fatima, ma tante Samira et toute la famille pour leurs soutiens et encouragements.

À mes chers amis, et en particulier à mon collègue Mehiris Hicham, je suis extrêmement reconnaissante pour leur soutien et leur présence tout au long de ce parcours.

Je souhaite exprimer ma gratitude particulière à ma belle Amira, dont la présence dans ma vie est une source d'appréciation infinie.

Enfin, je tiens à remercier du fond du cœur ma collègue Houda, avec qui j'ai partagé les moments les plus précieux tout au long de la réalisation de ce travail.

Sahra

Remerciements

En premier lieu, je rends grâce à Dieu, le Tout-Puissant à qui j'exprime ma vgratitude pour m'avoir donné, le courage et la patience et la volonté de réaliser ce travail.

Je tiens à remercier sincèrement Mr Nouioua Mourad pour son encadrement et son soutien et ses conseils précieux tout au long de ma recherche.

Mes remerciements vont également aux membres de jury pour l'intérêt qu'ils ont porté à mon projet en acceptant D'examiner mon travail et de l'enrichir de leurs propositions.

Évidemment, je n'oublie pas de remercier ma famille, mes amis et mes collègues et tous ceux qui, avec leur aide, leurs conseils, leurs encouragements, m'ont aidé à mener ce travail à terme

Résumé

Détecter la maladie avant qu'elle ne survienne est l'un des facteurs les plus importants du traitement médical. Ces dernières années, le domaine médical a connu une énorme expansion dans le domaine de l'informatique, comme l'apprentissage automatique et l'apprentissage en profondeur, ces modernes techniques ont été largement utilisés pour détecter diverses maladies. Cette procédure préventive est essentielle pour traiter les maladies à un stade précoce avant qu'elles ne se développent en maladies plus dévastatrices.

L'objectif de notre projet est la détection de maladies à l'aide de méthodes d'apprentissage automatique supervisé. Pour ce faire, nous avons utilisé quatre algorithmes de classification supervisée : Machine à Vecteurs de Supports (SVM), Arbres de décisions (DT), K-plus proche voisins (KNN) et La régression logistique (LR), pour trouver celui avec le taux de réussite le plus élevé.

Les algorithmes sélectionnés sont utilisés pour la prédiction de deux maladies : le diabète et l'insuffisance cardiaque. Les résultats obtenus prouvent l'efficacité de nos algorithmes améliorés étudies (panélisées). Spécialement, KNN qui a obtenu la meilleure performance.

Mots-clés : Apprentissage automatique, prédiction, classification.

Abstract

Detecting disease before it occurs is one of the most important factors in medical treatment. In recent years, the medical field has seen a huge expansion in the field of computer science, such as machine learning and deep learning, these modern techniques have been widely used to detect various diseases. This preventive procedure is essential to treat diseases at an early stage before they develop into more devastating diseases.

The objective of our project is the detection of diseases using supervised machine learning methods. To do this, we used four supervised classification algorithms: Support Vector Machine (SVM), Decision Tree (DT), K-nearest neighbors (KNN) and Logistic regression (LR), to find the one with the highest performance.

The selected algorithms are used for the prediction of two diseases: diabetes and heart disease. The obtained results prove the efficiency of our improved algorithms. Specially, KNN who got the best performance.

Keywords: machine learning, prediction, classification.

ملخص

يعد اكتشاف المرض قبل حدوثه من أهم العوامل في العلاج الطبي. في السنوات الأخيرة، شهد المجال الطبي توسعًا هائلاً في مجال علوم الكمبيوتر، مثل التعلم الآلي والتعلم العميق، وقد تم استخدام هذه التقنيات الحديثة على نطاق واسع للكشف عن الأمراض المختلفة. هذا الإجراء الوقائي ضروري لعلاج الأمراض في مرحلة مبكرة قبل أن تتطور إلى أمراض أكثر تدميراً.

الهدف من مشروعنا هو الكشف عن الأمراض باستخدام طرق التعلم الآلي الخاضعة للإشراف. للقيام بذلك، استخدمنا أربع خوارزميات تصنيف خاضعة للإشراف: دعم آلة المتجهات (SVM) ، شجرة القرار (DT) ، خوارزمية أقرب جيران (KNN) K والانحدار اللوجستي (LR) ، للتعرف على أحسن خوارزمي من حيث الأداء.

استخدمت الخوارزميات المختارة للتنبؤ بمرضين: مرض السكري وأمراض القلب. النتائج التي تم الحصول عليها تثبت كفاءة خوارزمياتنا المحسنة. على وجه الخصوص، KNN الذي حصل على أفضل أداء.

الكلمات المفتاحية: التعلم الآلي، التنبؤ، التصنيف.

Table des matières

Liste des abréviations

ML : Machine Learning

IA : Intelligence artificielle

LR: Logistic Regression

DT: Decision Tree

RL: Régréssion Logistique

KNN: K- Nearest Neighbour

SVM: Support Vector Machine

TP: Ture Positive

FP: False Negative

TN: True Negative

FN : False Negative

Liste des figures

Liste des tableaux

Introduction Générale et Problématique

- **Contexte Général :**

La santé est l'efficacité fonctionnelle et métabolique du corps et sa capacité à s'adapter aux changements physiques, mentaux et sociaux auxquels il est exposé. Les personnes psychologiquement saines se sentent à l'aise et heureuses dans leur vie et profitent bien de la vie. Une personne qui souffre de maladies mentales la voit toujours comme une personne sombre et pessimiste et ne se sent pas heureuse dans sa vie et dans ses relations avec les gens. [1]

L'importance de la santé pour une personne est qu'elle lui évite les frais de traitement et les tracas d'aller aux établissements de santé. Une personne en bonne santé qui est loin de la maladie la voit économiser de l'argent qu'elle aurait dépensé pour la maladie si elle avait négligé sa santé, alors on dit toujours qu'"une once de prévention vaut mieux que guérir". [2]

Une personne en bonne santé est une personne capable de se servir elle-même, sa nation et sa communauté. La société est également affectée par la présence de la maladie parmi ses membres, car la productivité de ces patients s'affaiblit et ils deviennent dépendants de la société. Quant à la présence d'individus en bonne santé dans la société, cela signifie la présence d'individus productifs qui possèdent le pouvoir et la capacité de donner et de servir. [2]

La santé prend forme sous nos yeux avec les progrès des technologies numériques de la santé telles que l'intelligence artificielle (IA), l'impression 3D, la robotique, la nanotechnologie, etc. La santé numérisée présente de nombreuses opportunités pour réduire les erreurs humaines, améliorer les résultats cliniques, suivre les données au fil du temps., etc. [3]

Les techniques d'intelligence artificielle (IA) se sont révélées extrêmement efficaces pour identifier et diagnostiquer différents types de maladies. Les chercheurs ont exploré une variété de techniques basées sur l'IA, notamment des modèles d'apprentissage automatique, afin de détecter et classifier les maladies.

Ces systèmes ne sont pas faits pour remplacer les spécialistes ou les médecins, mais ils sont développés pour aider les praticiens dans le diagnostic et la prévision de l'état du patient en se basant sur certaines règles ou expérience ou symptômes de maladies.

- **Problématique**

Des progrès nouveaux qui annoncent chaque jour, conduisent de plus en plus de médecins à s'interroger, à exprimer leur perplexité voire leurs craintes face à des évolutions pour lesquelles ils n'ont pas été préparés et qui paraissent leur échapper. [2]

Dans la pratique traditionnelle, le pronostic et la prédiction du risque de maladie reposaient sur la méthode statistique standard ainsi que sur l'intuition, les connaissances et l'expérience du médecin. Cependant, cette approche présente des risques de biais subjectifs et de variations individuelles dans l'évaluation du pronostic et du risque. Ces facteurs peuvent entraîner des erreurs de diagnostic, des décisions inappropriées et une qualité de soins inégale pour les patients.

La tâche de diagnostic est parfois très difficile à effectuer par les médecins et prendre de bonnes décisions à cause de : La grande quantité des données, la grande complexité des problèmes médicaux, la capacité du médecin est limitée pour résoudre quelques problèmes médicaux...etc.

Ces données ne peuvent pas être traitées par des méthodes classiques. Il faut donc utiliser de nouvelles techniques d'analyse pour réaliser des modèles afin de faciliter la prise de décision.

- **Motivation et Contributions**

Les avancées technologiques et les progrès de l'intelligence artificielle offrent de nouvelles opportunités pour améliorer la prise de décision médicale. En fait, les algorithmes de l'intelligence artificielle peuvent analyser de grandes quantités de données cliniques, génomiques et de santé, en identifiant des schémas, des tendances et des associations qui échappent souvent à l'œil humain. Cette capacité d'analyse approfondie peut aider les médecins à formuler des diagnostics plus précis, à prévoir les résultats des traitements et à élaborer des stratégies thérapeutiques plus efficaces. [2]

L’apprentissage automatique est une discipline essentielle de l'intelligence artificielle (IA). Il englobe un ensemble de techniques et d'algorithmes qui ont démontré leurs efficacités dans le domaine de la prédiction des maladies.

Dans le cadre de ce projet, notre objectif est de tirer parti des avancées récentes de l'apprentissage automatique afin de développer des modèles prédictifs capable d'assister les professionnels de la santé dans l'identification et la prédiction des maladies en se basant sur les symptômes.

En utilisant des données étiquetées qui associent les symptômes à des maladies spécifiques, les algorithmes d'apprentissage automatique supervisé seront en mesure d'apprendre des motifs et de faire des prédictions précises sur des données non observées. En exploitant ces modèles, nous espérons d’améliorer les performances diagnostiques, faciliter la prise de décision médicale et ouvrir de nouvelles perspectives pour une meilleure gestion des maladies.

Les contributions de ce travail peuvent être résumées comme suit

1. Nous utilisons les méthodes d’apprentissage supervisé pour développer quatre modèles de prédiction en se basant sur les méthodes suivantes : Machine à Vecteurs de Supports (SVM), Arbres de décisions (DT), K-plus proche voisins (KNN) et La régression logistique (LR). Les modèles ont été construits pour deux bases des maladies.
2. Nous développons une application graphique conviviale, dont le but de faciliter l'utilisation des modèles de prédiction des maladies proposés. Cette application fournit des interfaces graphiques qui permettent aux utilisateurs, tels que les professionnels de la santé, d'interagir facilement avec les modèles de prédiction.

- ## L’organisation du mémoire

Notre mémoire est organisé en quatre chapitres :

Dans le **première chapitre** nous offrons un aperçu concis de l'apprentissage automatique. Nous commençons par fournir une définition générale de l'apprentissage automatique, en soulignant son importance et son rôle. Ensuite, nous examinons en détail quatre types

d'apprentissage automatique, en fournissant leurs définitions, des exemples concrets, ainsi que les avantages et les inconvénients associés à chacun d'eux. Cette exploration approfondie nous permet de mieux comprendre les caractéristiques et les applications spécifiques de chaque type d'apprentissage automatique.

Dans le **deuxième chapitre**, nous proposons un aperçu concis de l'état de l'art de l'apprentissage automatique pour la prédiction des maladies basée sur les symptômes. Nous mettons en évidence l'importance du domaine de la santé et la nécessité d'utiliser des méthodes d'apprentissage automatique dans ce contexte. Nous explorons également les différentes applications des méthodes d'apprentissage automatique dans le domaine de la médecine et des soins de santé, en mettant l'accent sur la prédiction et le diagnostic des maladies. Nous examinons les avancées récentes dans ce domaine et les approches utilisées pour améliorer la précision et l'efficacité des modèles de prédiction et de diagnostic des maladies.

Le **troisième chapitre** se concentre sur les détails de la conception de notre travail. Après avoir introduit un schéma récapitulatif des étapes de notre travail, nous présentons les bases de données utilisées dans Cette étude, en expliquant leur origine et leur contenu. Ensuite, nous consacrons le reste du chapitre aux méthodes de classification supervisée que nous avons utilisées, notamment l'arbre de décision, les K plus proches voisins (KNN), la régression logistique, les machines à vecteurs de support (SVM), ainsi que les outils matériels et logiciels que nous avons utilisés pour mettre en œuvre ces méthodes. Enfin, nous décrivons les mesures que nous avons utilisées pour évaluer la performance de nos modèles de classification, en expliquant leur utilité et leur interprétation et Résultats des différentes méthodes supervisées.

Enfin, dans le dernier chapitre, nous présentons notre application avec ses différentes interfaces, en décrivant en détail son développement et les fonctionnalités qu’elle offre aux utilisateurs. Nous expliquons comment notre application a été conçue pour répondre aux besoins spécifiques identifiés dans notre étude.

Ensuite, nous concluons notre mémoire en dressant un bilan de cette étude. Nous mettons en évidence les contributions de notre travail, en soulignant comment notre recherche a contribué à l’avancement des connaissances dans le domaine du Data Maning et de l’apprentissage automatique appliqués à la prédiction des maladies.

Chapitre 1 : Apprentissage Automatique

1.1 Introduction

L'apprentissage automatique, également connu sous le nom de *« machine learning »*, est un domaine de l'intelligence artificielle qui permet aux machines d'apprendre et d'évoluer à travers un processus d'apprentissage basé sur les données. Contrairement à la programmation traditionnelle, où les instructions sont explicitement fournies, l'apprentissage automatique permet à la machine de découvrir des modèles et des relations dans les données pour effectuer des tâches complexes.

Dans ce chapitre, nous nous intéressons aux techniques d'apprentissage automatique. Dans une première partie, nous présentons les deux types d'apprentissage automatique : l'apprentissage supervisé et l'apprentissage non supervisé. Ensuite, nous présentons les méthodes de classification supervisées qui ont été utilisées dans notre étude notamment, SVM, Logistique Régression, Arbres de décision et K-plus proche voisins.

1.2 L'apprentissage Automatique « Machin Learning »

L'intérêt de l'apprentissage automatique a augmenté au cours de la dernière décennie, pour tout le discours sur l'apprentissage automatique, il y a beaucoup de conflits entre ce que la machine peut faire et ce que nous souhaitons.

D'une façon générale, l'apprentissage automatique est un type d'intelligence artificielle (lA), c'est une science qui permet aux ordinateurs d'apprendre sans être explicitement programmés « Arthur Samuel, 1959 ». Plus précisément, 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.

Dans le domaine médicale, l'apprentissage automatique a été conçu pour réaliser l'analyse de données médicales, surtout lorsque l'évolution numérique a fourni des moyens (capteurs) peu coûteux permettant de recueillir et de stocker des informations importantes liées aux patients et maladies. Par exemple, les algorithmes d'apprentissage sont utiles au médecin lors du diagnostic des patients, afin d'améliorer la vitesse, la précision et la fiabilité de son diagnostic.

1.3 Types d'Apprentissage Automatique

Il existe fondamentalement trois types d'apprentissage automatique : Supervisé, semi-supervisé et non-supervisé. Dans notre étude, nous utilisons l'apprentissage supervisé pour construite des modèles pour la prédiction des maladies.

Dans la suite de cette section, nous allons présenter les deux types d'apprentissage les plus utilisés qui sont l'apprentissage supervisé et apprentissage non supervisé.

La donne une vue globale de ces deux types d'apprentissage automatique.

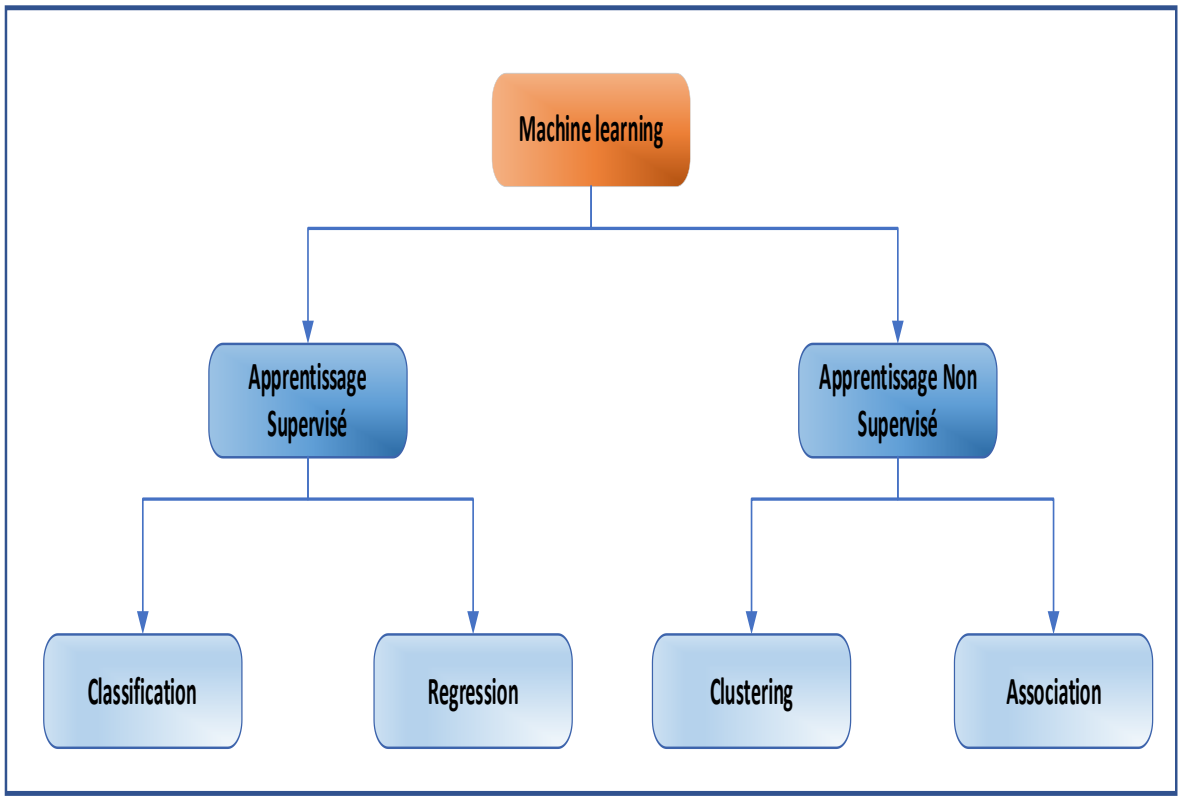


Figure 1-1: Les différentes méthodes d'apprentissage automatique.

1.3.1 Apprentissage Supervisé

L'apprentissage supervisé, en anglais « Supervised Learning », est un paradigme d'apprentissage automatique pour les problèmes où les données disponibles consistent en des exemples étiquetés, ce qui signifie que chaque point de données contient des caractéristiques (variables prédictifs) et une étiquette associée (variable cible).

L'objectif des algorithmes d'apprentissage supervisé est d'apprendre une fonction qui mappe les vecteurs de caractéristiques (entrées) aux étiquettes (sortie), sur la base d'exemples de paires entrée-sortie. Comme il est illustré dans la **Figure 1.1** l'apprentissage supervisé peut être utilisé pour deux types de tâches principales : la classification et la régression. Les algorithmes de classification cherchent à prédire la classe ou la catégorie à laquelle appartient une donnée d'entrée tandis que les algorithmes de régression servent à prédire une valeur numérique continue à partir de variables d'entrée. Dans ce travail, nous nous intéressons aux méthodes de classification.

1.3.2 Apprentissage Non-Supervisé

Contrairement à l'apprentissage supervisé, les méthodes d'apprentissage non supervisé doivent opérer à partir d'exemples non étiquetés. Ces méthodes doivent extraire automatiquement les catégories à associer aux données qu'on lui soumet, les plus fréquents problèmes connus dans ce type sont :

- Le clustering qui consiste à regrouper un ensemble d'éléments hétérogènes sous forme de sous-groupes homogènes.
- La réduction de dimension qui consiste à prendre des données dans un espace de grande dimension, et à les remplacer par des données dans un espace de plus petite dimension sans perdre la variance [4].

1.4 Algorithmes de Classification

Dans cette partie, nous présentons les algorithmes de classification utilisés dans notre travail. Nous décrivons le principe de leur fonctionnement, leurs avantages et leurs inconvénients. Les

définitions et explications de cette section sont tirées essentiellement des références suivantes : [5]

1.4.1 Les Arbres de Décision (Decision Trees)

Un arbre de décision, en anglais *« Decision Tree (DT) »*, est un schéma représentant les résultats possibles d'une série de choix interconnectés. Il permet à une personne ou une organisation d'évaluer différentes actions possibles en fonction de leur coût, leur probabilité et leurs bénéfices. Il peut être utilisé pour alimenter une discussion informelle ou pour générer un algorithme qui détermine le meilleur choix de façon mathématique.

Un arbre de décision commence généralement par un nœud d'où découlent plusieurs résultats possibles. Chacun de ces résultats mène à d'autres nœuds, d'où émanent d'autres possibilités. Le schéma ainsi obtenu rappelle la forme d'un arbre. Dans ces structures d'arbre, les feuilles représentent les valeurs de la variable-cible et les embranchements correspondent à des combinaisons de variables d'entrée qui mènent à ces valeurs.

1.4.1.1 Construction

- L’idée de construction de l’arbre de décision est simple : il faut commencer par diviser récursivement et le plus efficacement possible les exemples de l’ensemble d’apprentissage.
- Dans ces structures d'arbre, les feuilles représentent les valeurs de la variable-cible et les embranchements correspondent à des combinaisons de variables d'entrée qui mènent à ces valeurs.
- Principe de la construction : Au départ, les points de la base d’apprentissage sont tous placés dans le nœud racine. Une des variables de description des points est la classe du point, cette variable est dite « variable cible ».

- Chaque nœud est coupé (opération split) donnant naissance à plusieurs nœuds descendants. Un élément de la base d'apprentissage situé dans un nœud se retrouvera dans un seul de ses descendants.
- Le processus s'arrête quand les éléments d'un nœud ont la même valeur pour la variable cible.

1.4.1.2 Avantage des arbres de décision

1. Faciles à expliquer et comprendre.
2. Fonctionne avec des données catégorielles et numériques.
3. Peu couteux en termes de calcul.

1.4.1.3 Inconvénients des arbres de décision

1. Ils nécessitent souvent plus de temps pour former le modèle.
2. L'arbre devient plus complexe à mesure lorsqu'il s'approfondit.
3. Un petit changement dans les données peut entrainer un changement global de la structure de l'arbre de décision.

1.4.2 Machine à Vecteurs de Supports (Support Vector Machine)

Support Vector Machine (SVM) est l'un des algorithmes d'apprentissage supervisé les plus populaires, utilisé pour les problèmes de classification et de régression. Cependant, il est principalement utilisé pour les problèmes de classification dans l'apprentissage automatique. Le but de l'algorithme SVM est de créer la meilleure ligne ou limite de décision qui peut séparer l'espace à n dimensions en classes afin que nous puissions facilement mettre le nouveau point de données dans la bonne classe à l'avenir.

Cette meilleure frontière de décision est appelée un ***hyperplan***. SVM choisit les points - vecteurs extrêmes qui aident à créer l'hyperplan. Ces cas extrêmes sont appelés ***vecteurs de support***, et donc l'algorithme est appelé machine de vecteur de support**.** [6]

La figure suivante illustre deux classes différentes (classe des points bleus et classe des points roses) qui sont classés avec un hyperplan.

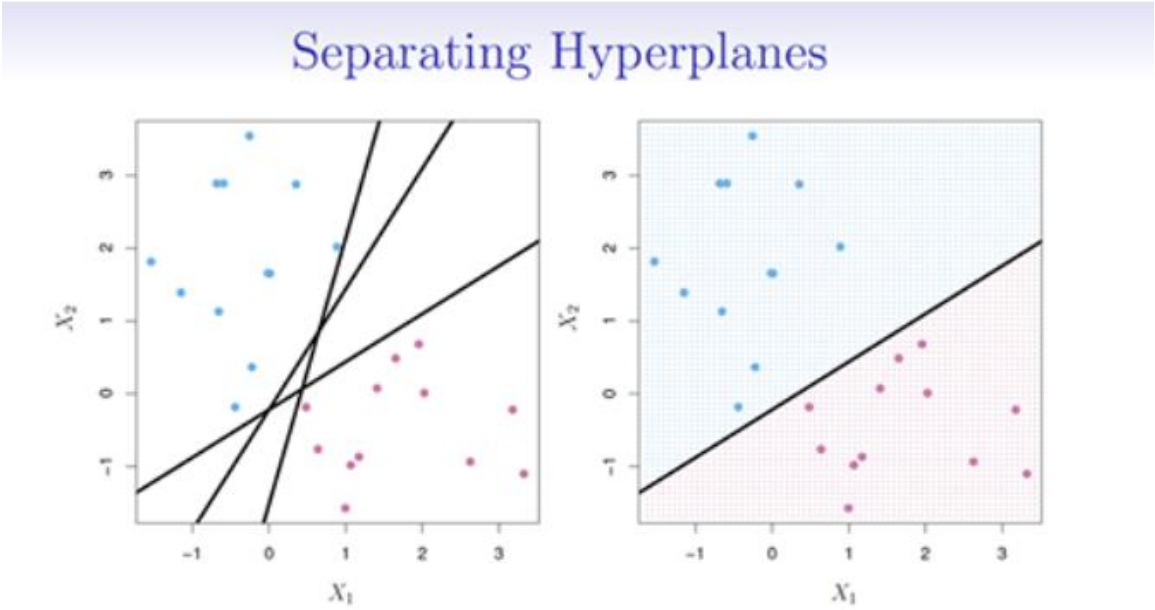


Figure1-2 : Séparation parfaite de deux classes avec un hyperplan.

1.4.2.1 Hyperplan, vecteur de support et marge dans l'algorithme SVM

La **Figure 1.3** présente les différents concepts liés à la méthode SVM. On peut voir que les hyperplans sont des limites de décision qui aident à classer les points de données dans un espace à n dimensions. La dimension de l'hyperplan dépend au nombre des entités dans le jeu de données. Si le nombre d'entité égale à 2, l'hyperplan sera une ligne. Et si le nombre d'entité égale à 3, l'hyperplan devient un plan bidimensionnel. [7]

Les vecteurs de support sont des points de données plus proches de l'hyperplan, et influencent la position et l'orientation de l'hyperplan. La position de l'hyperplan est dépanadant à la positon des vecteurs de support, La marge est la distance entre les vecteurs de support et l'hyperplan. [7]

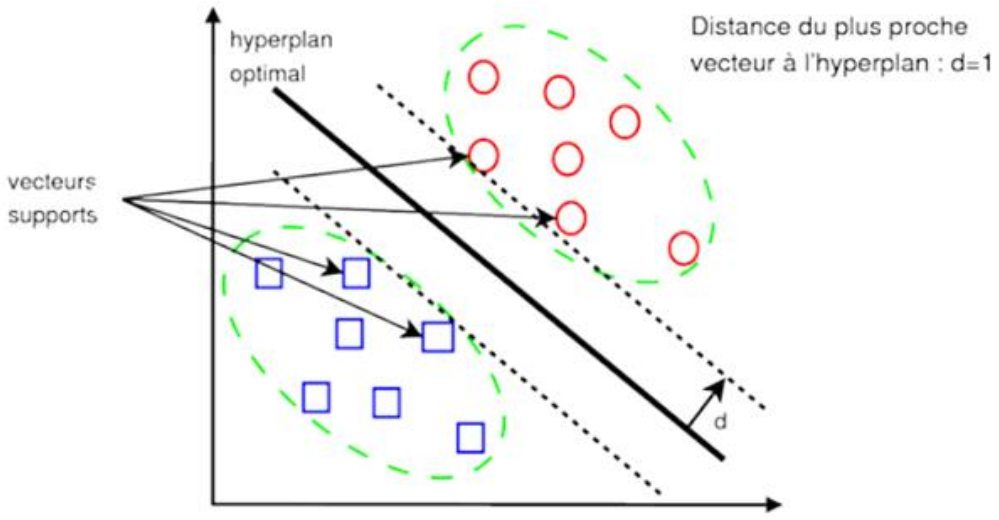


Figure 1-3: Les vecteurs de support, hyperplan et la marge.

1.4.2.2 Avantages de SVM

1. SVM a la capacité à gérer des frontières de décision non linéaires.
2. SVM fonctionne bien même avec des données non structurées et semi-structurées comme les textes, les images et les arbres.
3. SVM s'adapte relativement bien aux données de grande dimension.

1.4.2.3 Inconvénients de SVM

1. Complexité computationnelle : Les SVM peuvent être coûteux en termes de calculs, en particulier lorsqu'il s'agit de grands ensembles de données.
2. Difficile de comprendre et d'interpréter le modèle final, les poids variables et l'impact individuel.
3. L'extension de la classification à plus de deux classes est problématique.

1.4.3 Algorithme de K-Plus Proches Voisins (K-Nearest Neighbors)

K-plus proche voisins, en anglais *K-nearest neighbors (KNN),* est l'un des méthodes d'apprentissage supervisé le plus simple. Son fonctionnement est de classer les nouveaux points de données en fonction de la similarité aux points de données voisins.

Le principe de l'algorithme de k-plus proches voisins est le suivant : On suppose que l'ensemble *E* contient *n* données labellisées et *u* est une autre donnée qui n'appartenant pas à *E* qui ne possèdent pas de label. Soit *dist* une fonction qui renvoie la distance (qui reste à choisir) entre la donnée u et une donnée quelconque appartenant à E et soit *k* un entier inférieur ou égal à *n*. L'idée est la suivante :

1. Déterminer le paramètre *k* qui représente le nombre de plus proches voisins.
2. Calculer les distances entre la donnée *u* et chaque donnée appartenant à *E* à l'aide de la fonction de distance *dist*.
3. Considérons deux points$A = (x, y)$ et $B = (x', y')$ deux points du plan, la distance euclidienne entre *A* et *B* est :

$$dist(A, B) = AB = \sqrt{(x - x')^2 + (y - y')^2}$$

4. On retient les *k* données du jeu de données *E* les plus proches de *u*.
5. On attribue à *u* la classe qui est la plus fréquente parmi les *k* données les plus proches.

1.4.3.1 Exemple illustratif de KNN

La **Figure 1.4** présente un exemple de KNN méthode, dans cet exemple nous avons une donnée non classée (carré jaune), et toutes les autres données sont classées en une des deux classes : Classe A (étoile rouge) et classe B (triangle vert).

- Si *k=3*, les données les plus proche de la nouvelle donnée sont celles qui sont à l'intérieure du premier cercle. Dans ce cercle, la classe la plus prédominante est la classe B car on a 2 triangles verts et une seule étoile, donc la donnée non classée sera classée comme un triangle vert (classe B).
- Si *k=7*, les données les plus proches de la nouvelle donnée sont celle qui sont à l'intérieure du deuxième cercle. La classe la plus prédominante dans cette zone est l'étoile rouge (Classe A) car on a 4 étoiles rouges et 3 triangles, donc la donnée non classée sera classée comme une étoile rouge (Classe A). [8]

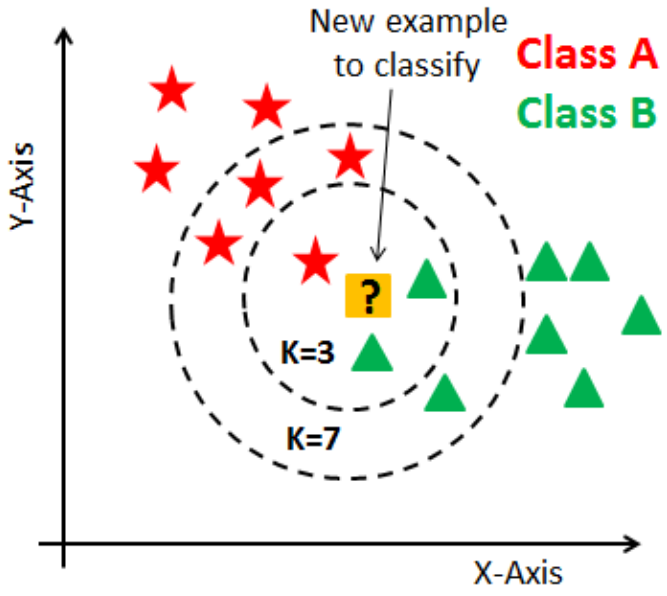


Figure 1-4 : Exemple illustratif de KNN.

1.4.3.2 Avantages de KNN

1. KNN est simple et facile à comprendre.
2. Interprétabilité : Le fonctionnement intuitif du KNN le rend facile à interpréter.
3. Peut être utilisée pour la classification et pour la régression.

1.4.3.3 Inconvénients de KNN

1. L'étape de prédiction peut être lente au cas où des grandes quantités de données.
2. Sensible à la dimensionnalité : KNN peut être moins efficace lorsque le nombre de dimensions (caractéristiques) est élevé.
3. Sensibilité à l'échelle des variables : Le KNN est sensible à l'échelle des variables.

1.4.4 Régression Logistique (Logistic Regression)

L'analyse de régression est souvent utilisée pour faire des prédictions, comprendre les variables indépendantes par rapport à la variable dépendante et étudier la forme de leur relation. Dans des circonstances limitées, l'analyse de régression peut être utilisée pour déduire la relation causale entre la variable indépendante et la variable dépendante. [9]

La régression logistique est un algorithme de classification qui est généralement mis en œuvre sur des problèmes de classification. Il est dérivé de la régression linéaire classique, où une fonction linéaire est utilisée pour prédire les résultats pour un ensemble d'entrées de données. La logistique.

La régression est un algorithme robuste lorsqu'il s'agit de classer des ensembles de problèmes, et a une fonction logistique (fonction sigmoïde) au cœur de celui-ci. Dans cet algorithme, les valeurs d'entrée sont combinées en fonction de coefficients ou de poids pour donner les valeurs de sortie/prédites. [10].

L'avantage de l'utilisation de la régression logistique par rapport à la régression linéaire est la capacité d'avoir une transition douce entre les prédictions des deux classes dans les problèmes de classification binaire. **La Figure 1.5** montre la différence entre le classifieur de la régression linéaire et du classifieur de ma régression logistique. [11]

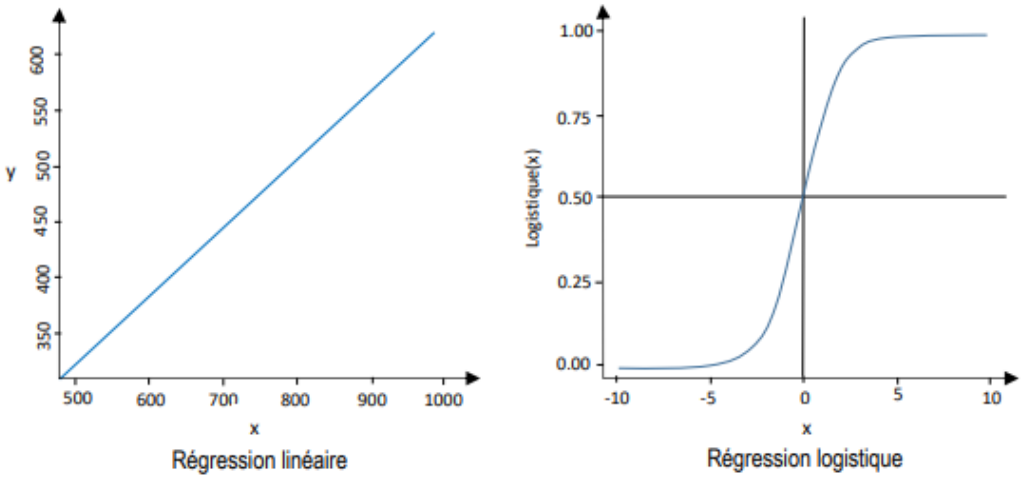


Figure 1-5: La différence entre la régression logistique et la régression linéaire.

1.4.4.1 Avantages de LR

1. La classification d'une nouvelle observation est extrêmement rapide puisqu'elle se résumé essentiellement à l'évaluation d'une fonction linéaire.
2. La simplicité de l'algorithme réduit sa susceptibilité au sur-apprentissage.
3. Ses résultats sont faciles à interpréter.

1.4.4.2 Inconvénient de LR

1. La phase d'apprentissage peut être longue car l'optimisation des coefficients est complexe.
2. Sa linéarité empêche la prise en compte des interactions entre les variables.
3. LR est principalement limité à des variables cibles binaires.

1.5 Conclusion

Dans ce chapitre, nous avons présenté les algorithmes d'apprentissage automatique. Après avoir présenté les deux principaux types d'apprentissage automatique, une description détaillée de chaque méthode de classification a été donnée en illustrant le principe de fonctionnement de chaque méthode ainsi que ses avantages et inconvénients. Le prochain chapitre présente quelques travaux connexes où les méthodes d'apprentissage automatique ont été appliquées dans le domaine de la santé.

Chapitre 02 : l'Apprentissage Automatique pour la Prédiction des Maladies basées sur les Symptômes.

2.1 Introduction

Les soins de santé, en anglais « HeathCare », sont un terme large qui concerne un système qui implique l'amélioration des services médicaux afin de répondre aux demandes médicales de la population. Dans le domaine de la santé, des efforts sont déployés par les patients, les médecins, les fournisseurs, les sociétés de santé et les sociétés informatiques pour maintenir et restaurer les dossiers de santé.

Au cours des dernières années, la recherche sur les soins de santé avec les méthodes d'apprentissage automatique (ML) n'a cessé d'augmenter. En raison de la variété des données médicales, y compris les données cliniques, les données omiques ou les données de santé électronique (DES), il est difficile pour les humains de déduire les données et de prendre des décisions. En conséquence, le ML a été proposé dans les soins de santé pour une meilleure compréhension des données et pour un meilleur processus de prise de décision.

Ce chapitre présente les diverses applications de ML méthodes dans le domaine de santé.

2.2 Healthcare et Apprentissage Automatique

La santé est une industrie avec forte évolution. En fait, nouvelles technologies et de nouveaux traitements sont développés en permanence, ce qui peut compliquer la tâche des professionnels de la santé. Pour pouvoir confronter ces complications, l'apprentissage automatique est devenu l'un des mots à la mode dans le domaine de la santé ces dernières années.

L'apprentissage automatique est particulièrement précieux car il peut nous aider à donner un sens aux énormes quantités de données, également les données de santé, qui sont générées chaque jour sous divers formats comme les dossiers de santé électroniques. L'utilisation des algorithmes d'apprentissage automatique peut nous aider à trouver des modèles et des connaissances qu'il serait impossible de les trouver manuellement. [12]

2.3Applications des ML pour la Médecine et Healthcare

Les techniques de ML peuvent servir à automatiser et à améliorer les performances dans les principaux secteurs applicatifs des soins de santé tels que le pronostic, le diagnostic, le traitement et le flux de travail clinique. Une représentation de la quantité de source de données hétérogènes dans les systèmes de santé est illustrée par la **Figure 2.1**. [13]

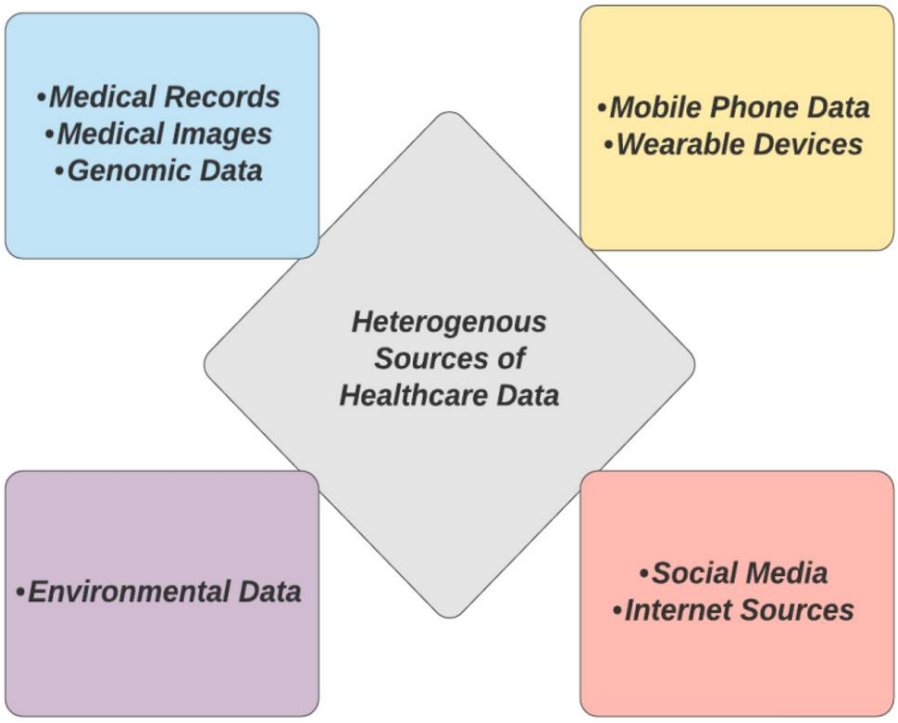


Figure 2-1: Illustration de sources hétérogènes contribuant aux systèmes de santé

2.3.1 Analyse des Dossiers de Santé Electroniques (Electronic Health Records)

Le dossier de santé électronique (DSE) est un dossier électronique longitudinal d'informations sur la santé des patients généré par une ou plusieurs rencontres dans n'importe quel cadre de prestation de soins. Ces informations comprennent les données démographiques des patients, les notes d'évolution, les problèmes, les médicaments, les signes vitaux, les antécédents médicaux, les vaccinations, les données de laboratoire et les rapports de radiologie.

Le DSE a la capacité de générer un dossier complet d'une rencontre clinique avec un patient, ainsi que de prendre en charge d'autres activités liées aux soins directement ou indirectement via l'interface, y compris l'aide à la décision fondée sur des preuves, la gestion de la qualité et la notification des résultats. [14]

Plusieurs méthodes d'apprentissage automatique ont été appliquées sur les dossiers de santé électroniques (DSE) : Dans le travail de Bricilmi Theodora et al, les auteurs visent à prédire les hospitalisations futures des patients souffrant de problèmes cardiaques ou de diabète, en fonction de leur DSE. Dans ce travail, le problème a été formulé comme un problème de classification binaire et trois méthodes de classification supervisée ont été adoptées : La méthode SVM, la méthode de régression logistique et la méthode de forêts aléatoires (Random Forests). Basé sur un objective similaire, des modèles prédictifs ont été proposé dans pour identifier les patients susceptibles d'être hospitalisés au cours de l'année suivante en raison de complications attribuées au diabète de type II. Les modèles proposés ont été testés sur des dossiers de santé électroniques extraite du centre médical de boston et les résultats obtenus ont démontré que les modèles peuvent peut prédire correctement environ de 81 % des patients hospitalisés. [15]

2.3.2 Prédiction et Diagnostique des maladies (Disease Prediction and Diagnosis)

Dans le domaine de la santé, la phase de diagnostic est essentielle pour l’orientation du patient et son suivi. L’apprentissage automatique apporte de nouvelles solutions aux professionnels de santé pour gagner du temps et faire le bon diagnostic. Plus précisément, les

méthodes prédictives d'apprentissage automatique incitent les moyens de pronostic et de diagnostic précoces à partir de données médicales qui réduisent le temps nécessaire pour agir sur la maladie pour le traitement. De plus, les méthodes d'apprentissage automatique ouvrent de nouvelles perspectives dans le repérage des maladies. Par exemple, il peut aider les médecins à détecter plus facilement les anomalies sur les radios des patients.

En fait, l'objectif n'est pas de remplacer le médecin par la machine, mais de l'accompagner dans l'analyse et l'interprétation des énormes volumes de données collectées. Machine Learning permet également de favoriser les bons diagnostics et de lutter contre les erreurs médicales en générant des diagnostics différentiels et suggérant des examens complémentaires.

L'identification et la prédiction de ces maladies à leurs premiers stades sont très importantes, afin d'en prévenir l'extrémité. Mais, la plupart du temps, il est difficile pour les médecins d'identifier manuellement les maladies avec précision. C'est lâ ou l'apprentissage automatique s'intervient pour nous garantir une fiable identification et prédiction des maladies.

Plusieurs travaux ont démontré l'efficacité des méthodes basé sur l'apprentissage automatique sur la détection et la prédiction des maladies, dans les auteurs ont illustré la capabilité des méthodes d'apprentissage automatique dans la prédiction de plusieurs types de maladies comme le cancer, les infections virulentes, la dengue, l'hépatite, les problèmes cardiaques, le paludisme, le diabète, etc. Akbulut et al. Ont proposé plusieurs techniques de ML pour surveiller et prédire l'état de santé du fœtus en fonction des antécédents cliniques de la mère. Durant la pandémique récente de COVID 19, les méthodes d'apprentissage automatique ont été aussi utilisées pour la prédiction et le diagnostic de la décompensation respiratoire chez les patients Covid-19.

Globalement parlant, l'apprentissage automatique avec ses déférentes formes a été successivement appliqué dans le domaine de prédiction des maladies. Cette réalité nous a motivé à utiliser les méthodes d'apprentissage automatique pour la prédiction des deux importantes maladies le diabète et les maladies cardiaques.

2.4 Autre Applications

Les techniques d'apprentissage automatique ont démontré leur succès dans autre domaines de soin de santé comme le traitement des images médicales. Les systèmes de ML ont enraciné leurs applicabilités dans les procédures d'analyse des images médicales. Ces techniques de calcul permettent l'extraction efficace d'informations importantes à partir d'échantillons d'images produits à l'aide de diverses modalités d'imagerie comme l'IRM, tomodensitométrie (CT), tomographie par émission de positrons (PET) et imagerie par ultrasons. Les progrès récents du matériel informatique permettent aux physiciens de réviser les anciens algorithmes d'IA et d'expérimenter de nouvelles idées mathématiques.

Un autre domaine de la santé qui a attiré l'attention des chercheurs de Machine Learning est le domaine de traitement médical. Le processus de médication suit une procédure de trois étapes : pronostic, diagnostic et traitement. Dans la phase de diagnostic, les images médicales sont étudiées par des cliniciens et des radiologues experts pour interpréter les risques et les remèdes possibles. Par conséquent, une grande quantité de données médicales est produite quotidiennement à partir de divers établissements de santé, les informations recueillies sont soumises à une supervision rigoureuse et les résultats sont consignés dans des rapports.

Cependant, la préparation de tels rapports nécessite une expertise et, si elle est gérée avec moins d'expérience dans les domaines des services de santé naissants, peut entraîner un diagnostic erroné ou peut se terminer par un synopsis critique. Par conséquent, les chercheurs ont tenté d'apporter des éclaircissements sur ces problèmes en utilisant les différentes techniques de ML.

2.5 Conclusion

Dans ce chapitre, nous avons présenté les domaines principaux où les méthodes d'apprentissage automatique ont été successivement appliquées. Parmi les divers domaines de santé, nous avons exposé l'application d'apprentissage automatique dans l'analyse des dossiers

de santé électroniques, l'analyse des images médicales, le traitement ainsi que la prédiction des maladies basées sur les symptômes qui est le sujet principal de notre étude.

Pour avoir plus des détails sur l'applicabilité de ML dans le domaine de la santé, le lecteur peut se référer aux articles suivants [16], dans lesquels des études détaillées ont été faites sur les diverses applications de machine Learning dans la santé.

Chapitre 03 : Conception et Réalisation

3.1 Introduction

Dans ce chapitre, nous débutons par la présentation des bases de données utilisées dans le cadre de ce projet, ainsi que les approches mises en œuvre pour atteindre notre objectif. Ensuite, nous présentons et nous comparons les résultats obtenus pour chaque base de données. Finalement, nous présentons notre application avec ses différentes interfaces.

3.2 Les bases de données

Dans notre étude, nous avons opté pour l'utilisation de deux bases de données basées sur les symptômes, avec un objectif de classification binaire. La première base de données concerne un groupe de femmes indiennes appartenant à la communauté *Pima*, vise à prédire si une personne est diabétique ou non[i].[25] Cette base de données comprend des informations telles que l'âge, le nombre de grossesses, la pression artérielle, le taux de glucose, etc. L'objectif est d'identifier les facteurs de risque de diabète chez cette population spécifique et de développer un modèle de prédiction précis. La deuxième base de données concerne les maladies cardiaques et contient plusieurs variables relatives à chaque patient, extraites de son dossier médical ainsi que des informations sur ses symptômes tel que l'âge, le sexe, le taux de cholestérol, etc.

3.3 Langage et bibliothèques utilisées

-  ***Python :*** est un langage de script de haut niveau, structuré et open source. Il est multi-paradigme et multi-usage. Développé à l'origine par Guido van Rossum en 1989, il est, comme la plupart des applications et outils open source, maintenu par une équipe de développeurs un peu partout dans le monde. Conçu pour être orienté objet, il n'en dispose pas moins d'outils permettant de se livrer à la programmation fonctionnelle ou impérative

; c'est d'ailleurs une des raisons qui lui vaut son appellation de « langage agile ». Il est connu pour sa rapidité de développement (qualité propre aux langages interprétés), la grande quantité de modules fournis dans la distribution de base ainsi que le nombre d'interfaces disponibles avec des bibliothèques écrites en C, C++ ou Fortran. C'est aussi un langage de programmation le plus utilisé dans le domaine du Machine Learning, du Big Data et de la Data Science. Python dispose de plusieurs bibliothèques tel que « Panda, Agate, Numpy,.Etc. ».

- ***Jupyter notebook* :** *Jupyter Notebook* est un environnement de développement interactif largement utilisé pour l'analyse de données, le prototypage et le développement de modèles d'apprentissage automatique. Il permet de combiner du code, des textes explicatifs, des visualisations et des résultats en un seul document. [17]
- ***Pandas* :** *Pandas* est une bibliothèque écrite en langage *Python* permettant la manipulation et l'analyse des données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques et de séries temporelles. [18]
- ***NumPy* :** *NumPy* est une extension du langage de programmation Python, destinée pour manipuler des tableaux multidimensionnels**.**
- ***Seaborn* :** *Seaborn* est une bibliothèque de visualisation de données *Python* basée sur *matplotlib*. Il fournit une interface de haut niveau pour dessiner des graphiques statistiques attrayants et informatifs. [19]
- ***Scikit-learn:*** *Scikit-learn* est une bibliothèque libre *Python* destinée à l'apprentissage automatique. Elle est développée par de nombreux contributeurs notamment dans le monde académique par des instituts français d'enseignement supérieur et de recherche comme *Inria*. [20]
- ***Matplotlib* :** est une bibliothèque *Python* destinée à tracer et visualiser des données sous formes de graphiques. [21]
- **GridsearchCv :** est une fonction fournie par la bibliothèque *scikit-learn* en *Python*. *Elle* est utilisée pour le réglage des hyperparamètres des modèles d'apprentissage automatique. Le but de GridSearchCv est de rechercher de manière exhaustive dans un

ensemble spécifié d'hyperparamètres et d'évaluer les performances du modèle à l'aide de la validation croisée pour déterminer la combinaison optimale d'hyperparamètres.

3.4 Mesures utilisées pour l'évaluation des modèles

Une matrice de confusion est un outil utilisé en apprentissage automatique et en statistiques pour évaluer les performances d'un modèle de classification. Elle est généralement utilisée pour comparer les prédictions d'un modèle avec les véritables valeurs cibles (ou étiquettes) des données.

C'est une matrice carrée qui présente les résultats des prédictions dans un format tabulaire. Elle comporte généralement deux dimensions : les classes réelles (ou étiquettes réelles) et les classes prédites (ou étiquettes prédites). Chaque case de la matrice représente le nombre (ou la fréquence) d'observations qui appartiennent simultanément à une certaine classe réelle et à une certaine classe prédite. Si on prend l'exemple de la prédiction des maladies, on obtient la matrice suivante :

		Réalité	
		N'est pas malade	Est malade
Prediction	N'est pas malade	Vrai négative (VN)	Faux positive (FP)
	Est malade	Faux negative (FN)	Vrai positive (VP)

Tableau 1:Un exemple d'une matrice de confusion.

La matrice de confusion est basée sur les valeurs suivantes :

- **VP : le nombre de vrais positifs (True positives).** VP représente le nombre d'instances qui sont correctement classées comme positive. En d'autres termes, VP est le nombre de fois où les valeurs réelles et prédites sont identiques et positives.
- **VN : le nombre de vrais négatifs (True positives*).*** VN représente le nombre d'instances qui sont correctement classées comme négative. En d'autres termes, VN est le nombre de fois où les valeurs réelles et prédites sont identiques et négatives.
- **FP : le nombre de faux positifs (False positives*).*** FP est le nombre d'instances incorrectement classés dans la classe positive. Dans notre cas, FP est le nombre de fois où le modèle prédit qu'un patient est malade, alors que le patient n'est pas malade en réalité.
- **FN : le nombre de faux négatifs (False negatives*):*** FN est le nombre d'instances incorrectement classés dans la classe négative. Dans notre cas, FN est le nombre de fois où le modèle prédit qu'un patient n'est pas malade, alors que le patient est malade en réalité.

A partir de ces valeurs, on peut définir les mesures retenues pour comparer et valider les différents modèles de classification :

- **Exactitude (Accuracy) :** La précision est l'une des principales mesures de performance pour la classification. Cette mesure représente le pourcentage de prédictions correctes effectuées par le modèle parmi l'ensemble des échantillons de données. L'exactitude est calculée par la formule suivante :

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

- **Rappel :** C'est la petite proportion des individus par rapport à la quantité globale des individus applicables. L'équation de rappel est représentée comme suit :

$$Rappel = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FN}}$$

- **Précision :** C’est la proportion des individus qui sont correctement identifiées par le modèle. L’équation de précision est représentée comme suit :

$$Précision = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FP}}$$

- **F1-score :** C'est la moyenne entre la précision et le rappel :

$$F1 - score = \frac{2 \times \mathrm{Précision} \times \mathrm{Rappel}}{Précision + Rappel}$$

3.5 Base de données du diabète

L’ensemble de donnée *Pima*, également connue sous le nom de "Pima Indians Diabetes Dataset", provient à l'origine de l'Institut national du diabète et des maladies digestives et rénales des Etats-Unis. L'objectif de l'ensemble de données est de prédire de manière diagnostique si un patient est diabétique ou non [22], sur la base de certaines mesures diagnostiques incluses dans l'ensemble de données. Plusieurs contraintes ont été placées sur la sélection de ces instances à partir d'une plus grande base de données. En particulier, tous les patients ici sont des femmes d'au moins 21 ans d'origine indienne.

L’ensemble des données se composent de plusieurs variables prédictives médicales et d'une variable cible (**Outcome**).

Les variables prédictives sont les suivantes :

- **Pregnancies:** Le nombre de grossesse.

- **Glucose :** La concentration plasmatique de glucose à 2 heures lors d'un test oral de tolérance au glucose.
- **Blood Pressure :** La pression artérielle diastolique (mm Hg).
- **Skin Thickness :** L'épaisseur du pli cutané du triceps (mm).
- **Insuline :** Insuline sérique de 2 heures (mu U/ml).
- **BMI:** Indice de masse corporelle (poids en kg/(taille en m)^2).
- **Diabetes Pedigree Function:** Fonction d'arbre généalogique du diabète.
- **Âge :** L'âge en années.

Le variable cible est outcome :

- **Outcome**: 1 signifie que le patient est malade et 0 signifie que le patient n'est pas malade.

3.5.1 Chargement des données

La première phase consiste à charger la base de données dans l'environnement jupyter notebook. La **figure 3.1** suivante donne le résultat de chargement l de la base *Pima*.

:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31.0	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0.0	0.232	54	1

Figure 3-1: Aperçu de l'ensemble de données Pima.

3.5.2 Prétraitement de données

L'étape de prétraitement des données est effectivement essentielle dans le domaine de l'apprentissage automatique (ML). Son objectif principal est de transformer l'ensemble des données brutes en un format approprié pour l'application des algorithmes de ML. Les données brutes peuvent être déformées, peu fiables ou contenir des valeurs manquantes, ce qui peut entraîner des problèmes lors de l'entraînement des modèles de ML.

3.5.3 Description de la base

Pour les données numériques, nous avons calculé différents indices de résultat. Ces indices comprennent le nombre d'échantillons, la moyenne, la norme, la valeur minimale (min), la valeur maximale (max) ainsi que les centiles inférieur, 50 et supérieur. Les résultats sont présentés dans la **Figure 3.2**. On peut voir que les variables prédictives ont des échelles différentes. Cela peut résulter que, certaines variables peuvent dominer les autres en termes de magnitude. Ce qui va conduire à des biais dans les modèles d'apprentissage automatique, où les variables avec des valeurs plus grandes pourraient avoir un impact disproportionné sur les résultats.

Pour éviter ce problème, une étape de mise à l'échelle des variables *(features scaling)* est nécessaire pour équilibrer les variables en les ramenant à une échelle comparable. Cette étape sera faite après la division des données.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	121.686763	72.405184	29.108073	140.671875	32.455208	0.471876	33.240885	0.348958
std	3.369578	30.435949	12.096346	8.791221	86.383060	6.875177	0.331329	11.760232	0.476951
min	0.000000	44.000000	24.000000	7.000000	14.000000	18.200000	0.078000	21.000000	0.000000
25%	1.000000	99.750000	64.000000	25.000000	121.500000	27.500000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.202592	29.000000	125.000000	32.300000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Figure 3-2:Description de la base Pima.

3.5.4 Exploration des données

Avant de mettre en œuvre les méthodes de classification, nous avons programmé des codes afin de visualiser les données et faciliter sa compréhension.

La **Figure 3.3** donne un graphe qui représente les nombre des femmes maladies et saines avec leurs pourcentages. On remarque que, le nombre de femmes diabétiques représente 34.9% du nombre total de patients alors que le nombre de femmes non diabétiques représente 65.1% du nombre total de patients.

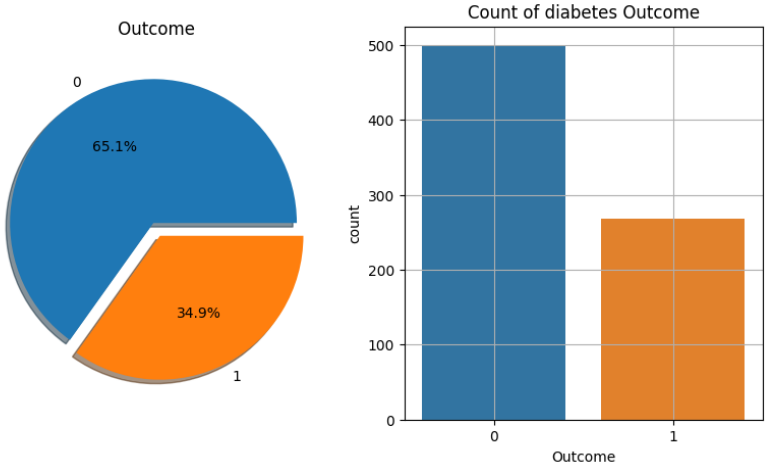


Figure 3-3: Nombre et Pourcentage des femmes maladies et saines.

La **Figure 3.4** représente la variation du nombre des femmes diabétiques et non diabétiques en fonction de nombre de grossesses. Au début avec un nombre de grossesse inférieur ou égale à 6, on peut voir que le nombre de femme diabétiques est toujours inférieur au nombre de femme non diabétiques. Après 6 grossesses, ce résultat est inversé et le nombre de femmes diabétiques devient plus grand que les nombre des femmes saines. Ce résultat signifie que dans cet échantillon de données, la maladie de diabète a une forte corrélation avec le nombre de grossesse.

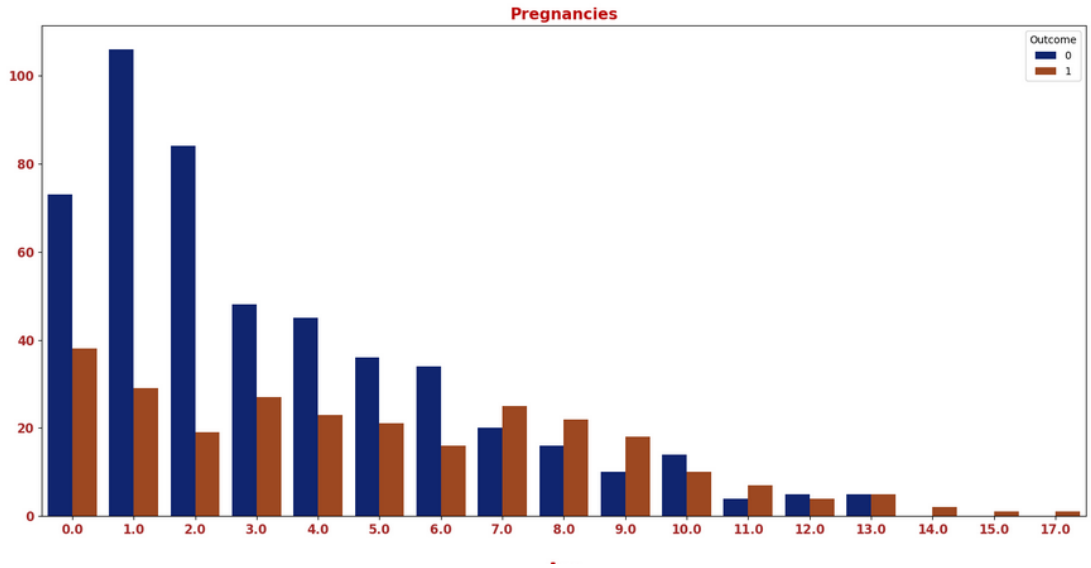


Figure 3-4:Fréquences du diabète par rapport au nombre de grossesse.

3.5.5 Division des données :

A ce point, l'ensemble de données est divisé en deux parties : un échantillon d'apprentissage (70% de donnéés) et un échantillon de test (30%). La fonction train_test_split de Scikit-learn est utilisée pour effectuer cette opération.

```
X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.3,shuffle="true",random_state=7)
```

```
print("X_test:",X_test.shape[0])
print("X_train:",X_train.shape[0])
```

```
X_test: 231
X_train: 537
```

Figure 3-5: Code division des données.

3.5.6 Feature Scaling

Comme il a été mentionné précédemment, une phase de feature scaling est nécessaire pour rendre les échelles des variables comparables et éviter la dominance des variables ayant les grandes échelles. Pour effectuer cette étape de normalisation, nous avons utilisé la méthode de MinMaxScaler. La méthode MinMaxScaler est appliquée sur l'échantillon d'apprentissage. Ensuite, une transformation de données est faite sur les deux échantillons.

3.5.7 Construction des modèles de classification

Après avoir transformé les données. Cette phase consiste à effectuer la classification supervisée en utilisant les méthodes KNN, SVM, LR et DT. Les modèles de classification sont construits à partir d'une phase d'apprentissage. L'objectif de la phase d'apprentissage est de construire un modèle prédictif à partir de l'échantillon d'apprentissage. En se basant sur les exemples donnés dans l'échantillon d'apprentissage, le modèle tente d'expliquer comment la variable cible est dépendante des variables prédictives. Le modèle peut alors prédire la variable cible pour un nouveau cas où seules les variables prédictives sont données.

3.5.8 Résultats et évaluations des différents modèles

Après avoir entraîné les modèles avec les algorithmes KNN, SVM, LR et DT, nous prédisons les classes en utilisant l'ensemble de données de test afin de sélectionner le meilleur modèle à utiliser. Le meilleur modèle est choisi en fonction de ses performances.

Comme il a été mentionné précédemment, les performances d'un algorithme de machine learning sont évaluées en fonction de paramètres tels que l'exactitude, la précision, rappel et F-1 score, etc. Dans notre étude, toutes ses mesures sont utilisées comme des critères de performance pour évaluer les différents modèles de classification.

Les matrices de confusion des différents modèles sont présentées dans le tableau suivant :

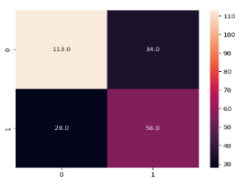
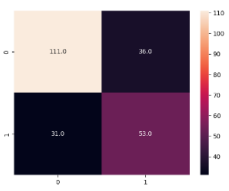
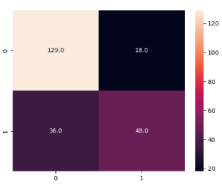
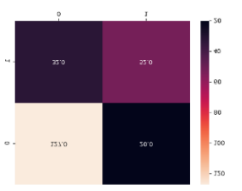
KNN			**DT**			**LR**			**SVM**		
	False	**True**		**False**	**True**		**False**	**True**		**False**	**True**
False	56	34	**False**	53	36	**False**	48	18	**False**	52	20
True	28	113	**True**	31	111	**True**	36	129	**True**	32	127
											

Tableau -2: Les matrices de confusion des différents modèles.

A partir de ces matrices de confusion, les métriques d'évaluation sont présentes dans la **Figure 3.6.**

Model	Training Accuracy	Testing Accuracy	Precision	Recall	F1-Score
Logistique regression	77.65%	76.62%	76.20%	76.62%	75.90%
Support Vector M(SVM)	77.65%	76.62%	77.09%	77.49%	77.06%
K-Neighbors (KNN)	77.65%	76.62%	73.63%	73.16%	73.34%
Decision Tree	77.65%	76.62%	71.40%	71.00%	71.16%

Figure 3-6: Les résultats de classification de différents algorithmes.

D'après les résultats de cette figure, nous voyons que, les quatre modèles ont la même valeur d'exactitude qui égale à 76,62 %. En termes de précision, rappel et f1-score, on voie que le SVM a obtenu le meilleur résultat suivi de LR, KNN et finalement DT.

3.5.9 Optimisation des hyperparamètres de différentes méthodes

Cette étape consiste à améliorer les résultats de différentes méthodes de classification. L'optimisation des hyperparamètres, également connue sous le nom de « hyperparameters tuning », est le processus visant à trouver la meilleure combinaison de valeurs d'hyperparamètres pour un algorithme d'apprentissage automatique. Les hyperparamètres sont des paramètres de configuration qui ne sont pas appris à partir des données, mais qui sont définis avant le début du processus d'apprentissage. Des exemples d'hyperparamètres incluent le nombre de voisins de KNN, le nombre de couches cachées dans un réseau neuronal, etc.

La figure ci-dessous montre que le nombre de voisins **k=15** donne le meilleur résultat d'exactitude avec 80.52 %. Ce résultat est largement meilleur que le résultat initial de KNN qui était juste 76.62%.

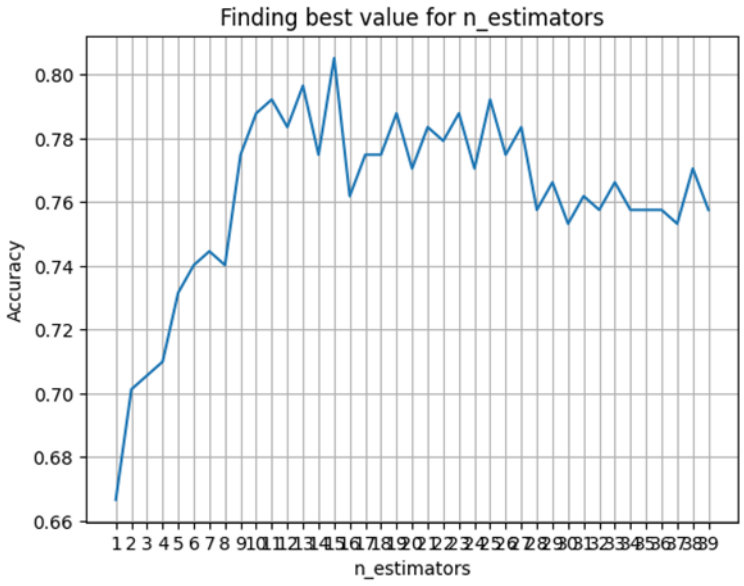


Figure 3-7: Le nombre optimal de k-voisins de KNN.

Le même processus d'optimisation a été appliqué pour les autres classificateurs. Les résultats sont présentés dans **Tableau -3** et **Figure 3.8**.

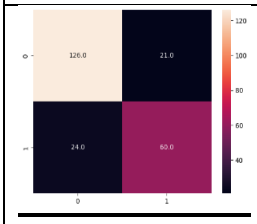
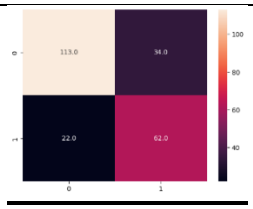
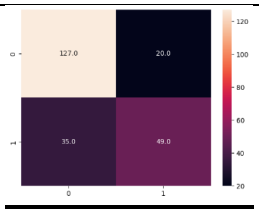
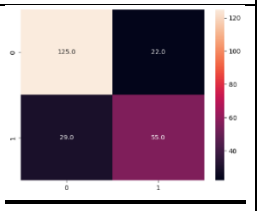
KNN			**DT**			**LR**			**SVM**		
	False	**True**		**False**	**True**		**False**	**True**		**False**	**True**
False	60	21	**False**	62	34	**False**	49	20	**False**	55	22
True	24	126	**True**	22	113	**True**	35	128	**True**	29	125
											

Tableau -3:Les matrices de confusion des différents modèles après l'optimisation des hyperparamètres

Model	Training Accuracy	Testing Accuracy	Precision	Recall	F1-Score
Logistique regrission	77.65%	76.19%	75.71%	76.19%	75.60%
Support Vector M(SVM)	82.31%	77.92%	77.63%	77.92%	77.70%
K-Neighbors (KNN)	78.96%	80.52%	80.39%	80.52%	80.44%
Decision Tree	86.78%	76.19%	77.07%	76.19%	76.46%

Figure 3-8: Les résultats de classification de différents algorithmes après l'optimisation des hyperparamètres.

Nous voyons que, la meilleure classification est obtenue par la méthode "KNN" avec une valeur d'exactitude de 80.52 %. Suivie par "SVM" avec 77,92%. Ensuite, les algorithmes LR et DT qui ont obtenu le même taux de prédiction de 76,19%.

3.6 Base de données des maladies cardiaques

La deuxième base de données utilisée dans ce travail est la base de données Heart Failure Prédiction Dataset [23]. Cette base contient des informations sur chaque patient, basées sur son dossier médical et ses symptômes. Cette base de données comprend 918 enregistrements et 12 attributs. Elle a été largement utilisée par les chercheurs en apprentissage automatique jusqu'à ce jour.

Le champ "HeartDisease" (maladie cardiaque) fait référence à la présence ou non d'une maladie cardiaque chez le patient. Les étiquettes de classe se composent de deux valeurs : 0 pour les personnes normales et 1 pour les patients atteints d'une maladie cardiaque.

Pour parvenir à un résultat permettant de déterminer si un patient est atteint d'une maladie cardiaque ou non, l'étude a été réalisée en utilisant 12 variables explicatives :

- **Age :** Age lors de la crise cardiaque.
- **Sex :** Sexe de la personne.
- **ChestPainType :** Type de douleur à la poitrine.

- **RestingBP :** Pression artérielle au repos (mm Hg).
- **Cholesterol :** Cholesterol du patient (mm/dl).
- **FastingBS :** Glycémie à jeun du patient (1 si FastingBS 120 mg/dl, 0 sinon).
- **RestingECG :** Résultat des électrocardiogrammes au repos (Normal : Normale, ST : Ondes ST-Tabnormales (Invesions onde T et/ou augmentaion ou réduction de > 0.05 mV des ondes ST), LVH : Hypertrophie probable ou définitive du ventricule gauche selon le critère d'Estes).
- **MaxHR :** Fréquence cardiaque maximale (entre 60 et 202 bpm).
- **ExerciseAngina :** Angine causée par l'exercice (Y : Oui, N : Non).
- **Oldpeak:** oldpeak = ST [Numeric value measured in depression].
- **ST_Slope:** La pente du segment ST d'effort maximal [Up: upsloping, Flat: flat, Dowdownsloping].

La variable cible est **HeartDisease**

- **HeartDisease *:*** Si le patient avait une maladie cardiaque (1: Oui, 0: Non).

3.6.1Chargement des données

La **Figure 3.9** donne une description de la base donnée. On remarque que contrairement à la base de diabètes, cette base de données contient aussi des variables qualitatives tels que : ExerciseAngina at ST_Slope. Pour cette raison, il est nécessaire de transformer ces variables qualitatives dans la phase de prétraitement pour permettre aux algorithmes d'apprentissage automatique de traiter et d'analyser ces variables.

```
data=pd.read_csv('heart.xls')
data.head()
```

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0

Figure 3-9:Aperçu de l'ensemble de données.

3.6.2 Prétraitement de données

Les algorithmes de ML sont principalement conçus pour traiter des données numériques. Alors, cette phase consiste à convertir nos données qualitatives en données numériques en utilisant l'encodage one-hot.

L'encodage one-hot sert à encoder les variables qualitatives nominales, où il n'y a pas d'ordre spécifique entre les catégories, l'encodage one-hot est souvent utilisé. Cela consiste à créer de nouvelles variables binaires (0 ou 1) pour chaque catégorie de la variable d'origine. Chaque nouvelle variable représente la présence ou l'absence de cette catégorie dans l'observation. Par exemple, si vous avez une variable "couleur" avec les catégories "rouge", "vert" et "bleu", elle sera transformée en trois variables binaires distinctes : "rouge" (1 ou 0), "vert" (1 ou 0) et "bleu" (1 ou 0). L'encodage one-hot est fait grâce à la fonction python **get_dummies.**

3.6.3 Exploration des données

Similaire à la première base, nous effectuons une exploration des données. La **Figure 3.10** représente les nombre des personnes maladies et saines avec leur pourcentage. On voie que, 55.3% de patients souffrent d'une maladie cardiaque.

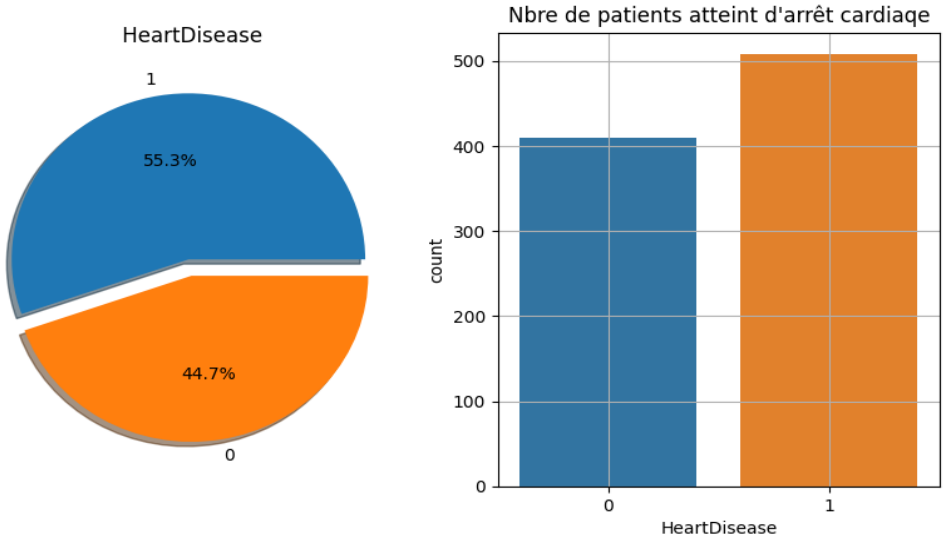


Figure 3-10: Nombre et pourcentage des personnes maladies et saines

La **Figure 3.11**représente le nombre d'homme et femmes dans la base de données.

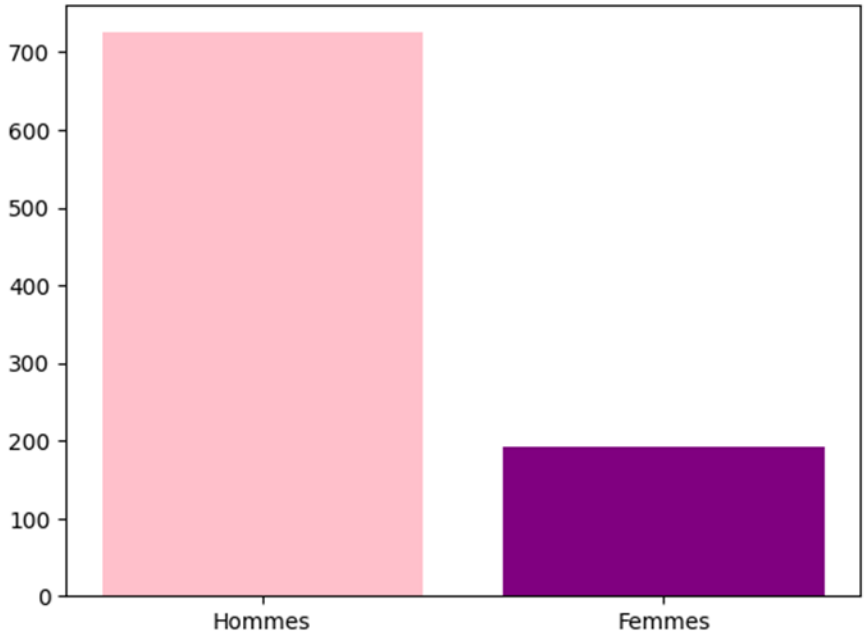


Figure 3-11:Nombre des femmes et hommes dans base de données.

3.6.4 Division des données

L'ensemble de données est divisé en ensembles d'apprentissage (70%) et de test (30%). En utilisant la fonction train_test_split.

3.6.5 Feature Scaling

Cette tâche consiste à transformer les variables quantitatives à fin pour rendre les échelles des variables comparables et éviter la dominance des variables ayant les grandes échelles.

```
scaler = MinMaxScaler()
scaler.fit(X_train)
```

```
▾ MinMaxScaler
MinMaxScaler()
```

```
X_train_scaled=scaler.transform(X_train)
X_test_scaled=scaler.transform(X_test)
```

Figure 3-12:Fréquence du diabète par rapport à Pregnancies

3.6.7 Construction des modèles de classification

Similaire à la base donnée de diabètes, après les étapes précédentes, nous effectuons la phase de construction des modèles de classification en utilisant les méthodes KNN, SVM, LR et DT.

3.6.8 Résultats et évaluations des différentes méthodes supervisées

Les résultats de classification des différentes méthodes sont présentés par **Tableau 4** et la **Figue 3.13.**

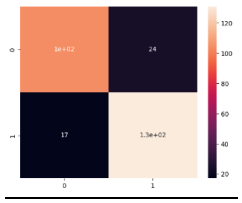
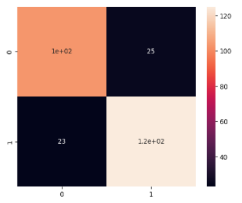
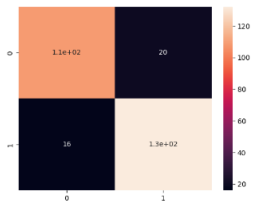
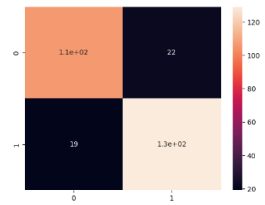
KNN			DT			LR			SVM		
	False	**True**		**False**	**True**		**False**	**True**		**False**	**True**
False	104	24	**False**	103	25	**False**	108	20	**False**	106	20
True	17	131	**True**	23	125	**True**	16	132	**True**	16	132
											

Tableau -4. Les matrices de confusion des différents modèles.

A partir des matrices de confusion, nous calculons les mesures d'évaluation de chaque algorithme. Les résultats sont présentés dans **Figue 3.13**.

Model	Training Accuracy	Testing Accuracy	Precision	Recall	F1-Score
Logistique regression	86.14%	86.96%	86.96%	86.96%	86.94%
Support Vector M(SVM)	86.14%	86.96%	85.14%	85.14%	85.13%
K-Neighbors (KNN)	86.14%	86.96%	85.18%	85.14%	85.11%
Decision Tree	86.14%	86.96%	82.60%	82.61%	82.60%

Figure 3-13:Les métriques de performance de chaque algorithme « Maladies Cardiaque ».

D'après les résultats présentés dans la table, nous observons que la méthode de logistique régression a obtenu le meilleur résultat, atteignant 86.96% en termes de précision et rappel et 86.94% en termes de F1-score. On voie aussi que, les résultats de SVM et KNN sont très compétitifs.

3.6.9 Optimisation des hyperparamètres de différentes méthodes

A ce point, une phase d'optimisation des hyperparamètres est effectuée pour améliorer les résultats des déférentes méthodes de classification.

Pour le modèle KNN, nous nous intéressons à trouver le nombre optimal de voisins. Le résultat obtenu est présenté par la **Figure 3.14.** Nous voyons que, la valeur de **K=71** a donné le meilleur résultat avec un taux d'exactitude de 87.32%.

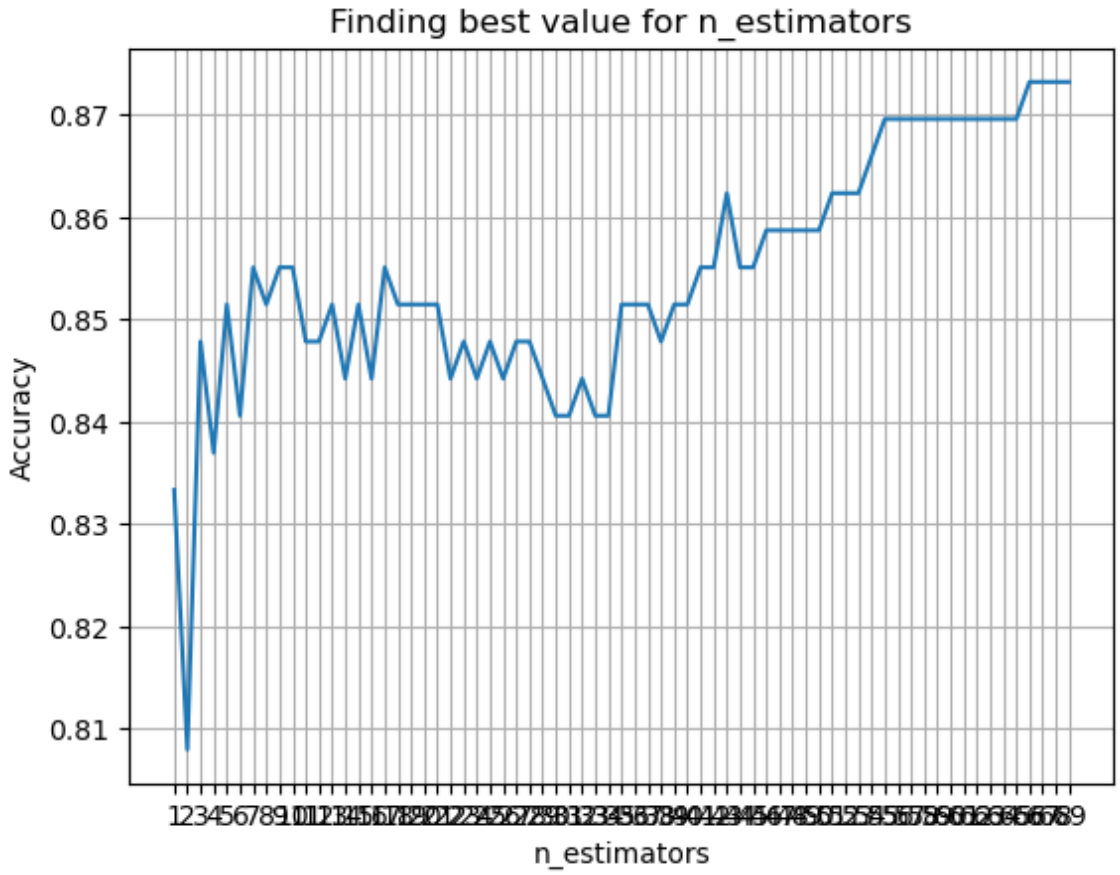


Figure 3-14: plot de déterminer meilleur K value KNN

Les résultats des différents modèles de classification après la phase d'optimisation des hyperparamètres sont présentés dans **Tableau -5** et **Figure 3.15**.

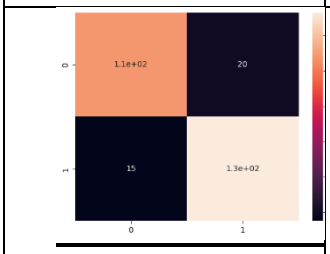
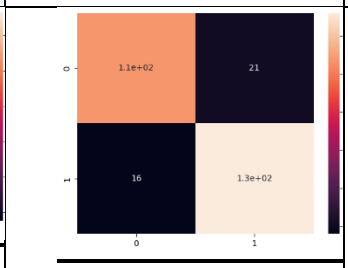
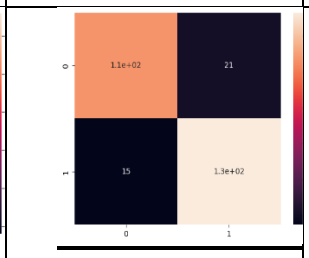
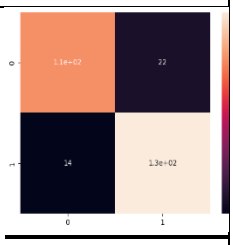
KNN			DT			LR			SVM		
	False	**True**		**False**	**True**		**False**	**True**		**False**	**True**
False	108	20	**False**	107	21	**False**	107	21	**False**	106	22
True	15	133	**True**	16	132	**True**	15	133	**True**	14	134
											

Tableau -5:la matrice de confusion des différents modèles.

Selon la figure ci-dessous, la meilleure classification est également obtenue par la méthode KNN avec un taux d'exactitude de 87,32 %. LR et SVM obtiennent la même valeur d'exactitude et ils ont des résultats compétitifs. Finalement, DT a obtenu des résultats moins optimaux que les autre classifieurs.

Model	Training Accuracy	Testing Accuracy	Precision	Recall	F1-Score
Logistique regression	87.07%	86.96%	86.99%	86.96%	86.93%
Support Vector M(SVM)	87.38%	86.96%	87.03%	86.96%	86.92%
K-Neighbors (KNN)	86.29%	87.32%	87.33%	87.32%	87.30%
Decision Tree	90.97%	86.96%	86.96%	86.96%	86.94%

Figure 3-15:Les métriques de performance de chaque algorithme « Maladies Cardiaques »

3.7 Discussion :

Les avancées d’e-santé, de la télémédecine et du traitement des données massives changent la manière de dispenser les soins de santé. Il est temps de faciliter accès à ces données massives en utilisant des systèmes informatisés, afin de permettre aux médecins et aux professionnels de la santé de diagnostiquer plus rapidement et précisément.

Dans ce contexte, notre travail est de proposer un outil d’aide à la décision pour diagnostiquer les maladies cardiaques et le diabète. Nous avons constaté que, la méthode KNN est un bon choix pour cela. Cette méthode ne remplace pas le médecin, mais elle constitue un outil intelligent pour exploiter les énormes quantités de données disponibles concernant des patients qui ont déjà fait le bon diagnostic.

3.8 Conclusion

Dans ce chapitre, nous avons présenté les différentes étapes que nous avons suivi pour développer et mettre en œuvre notre système de prédiction du diabète et des maladies cardiaques. Les résultats des performances des différents algorithmes sont ensuite présentés en détails en suivant diverses mesures de performance. Les résultats ont révélé que le KNN a obtenu la meilleure performance par rapport aux autres modèles de prédiction.

Chapitre 04 : Présentation de l'application

4.1 Introduction

Dans ce chapitre, nous allons passer à une phase importante de notre projet : L'implémentation de l'application. Premièrement, nous décrivons les outils et les langages de programmation utilisés, puis nous présentons l'application créée et nous illustrons comment elle peut être utilisée.

Le choix des outils de développement et des langages de programmation peuvent considérablement affecter le temps de programmation et la qualité du code produit. Cette phase vise à transformer les modèles de classification supervisée établis en une application destinée aux utilisateurs, en particulier les médecins. Pour cela, nous avons choisi d'utiliser *Python* comme langage de programmation et PyCharm comme environnement de développement.

4.2 Outils utilisés

- **PyCharm :** est un environnement de développement intégré utilisé pour programmer en Python. Il permet l'analyse de code et il contient un débogueur graphique. Il permet également la gestion des tests unitaires, l'intégration de logiciel de gestion de versions, et supporte le développement web avec Django. PyCharm est développé par l'entreprise Tchèque JetBrains, c'est un logiciel multi-plateforme qui fonctionne sous Windows, Mac OS et Linux. Il est décliné en édition professionnelle, diffusé sous licence propriétaire, et en édition communautaire diffusé sous licence Apache. [24]

- **Tkinter :** est une bibliothèque graphique pour *Python*. Elle permet de créer des interfaces graphiques utilisateur (GUI) pour des applications en utilisant le langage de programmation Python. Tkinter est inclus dans la bibliothèque standard de *Python*, ce qui

facilite son utilisation par les programmeurs Python. Tkinter fournit des widgets standard tels que des boutons, des étiquettes, des champs de saisie, des menus, des listes déroulantes, etc., qui peuvent être utilisés pour créer une interface utilisateur. Il est également possible de personnaliser l'apparence des widgets en utilisant des options de configuration. Tkinter offre également des fonctions pour la gestion d'événements, la gestion des boîtes de dialogue, la création de fenêtres, la mise en page des widgets, etc.

4.3 Mode d'utilisation de l'application

4.3.1 Interface « Menu principal »

L'illustration présentée dans la **Figure 4.1** montre la figure principale de l'application, communément appelé le menu principal. Cette fenêtre offre un menu qui permet de télécharger une base de données sur le diabète et les maladies cardiaques. Au sein de cette fenêtre, vous trouverez trois boutons : Le premier, nommé « **Diabète** », vous permet d'accéder à la base de données de diabètes Pima, tandis que le deuxième bouton, intitulé « **Maladies cardiaques** », vous donne accès à la base de données "Maladie cardiaque UCI". De plus, un bouton de fermeture est présent pour quitter l'application.

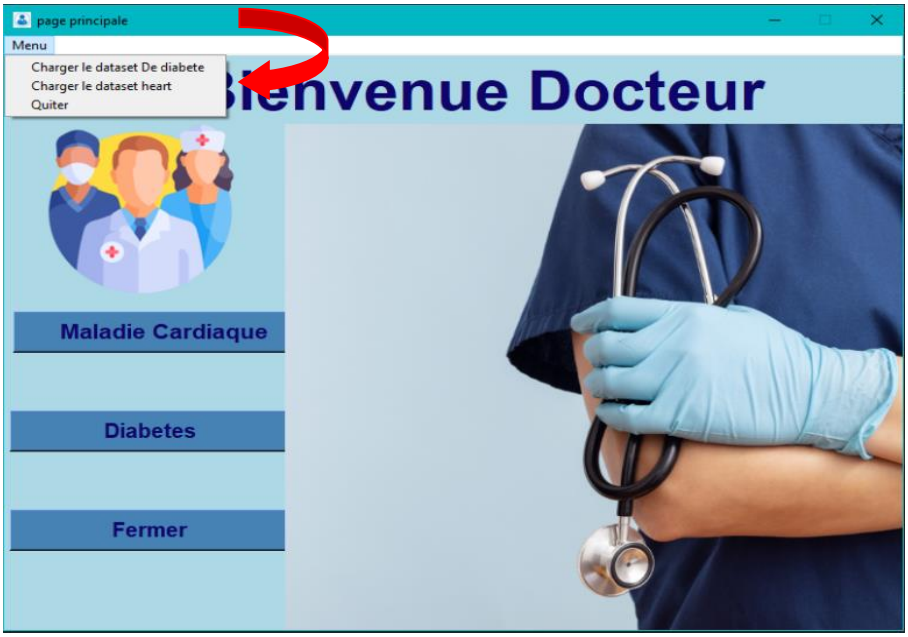


Figure 4-1: interface principale d'application

4.3.2 Interface-Diabète

En cliquant sur le bouton « **Diabète** », la **Figure 4.2** s'affiche en tant que première fenêtre. Elle comprend un menu déroulant permettant de sélectionner un modèle, ainsi que trois boutons.

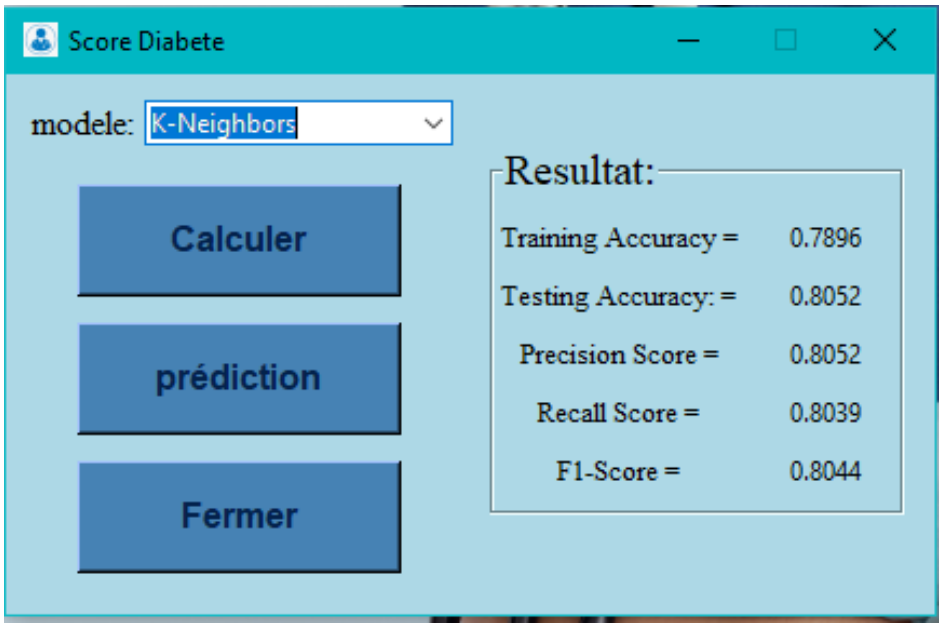


Figure 4-2:Précision d'apprentissage du diabète

Le premier bouton, nommé « **Calculer** », est utilisé pour afficher les résultats de classification d'un modèle choisis tel que : la précision, rappel et f1-score (voir **figure 4.2**).

Le bouton « **Prédiction** » sert à afficher la fenêtre de la **Figure 4.3** avec laquelle l'utilisateur peut remplir un formulaire qui contient les informations sur le patient dont nous voulons prédire son état (malade ou sain).

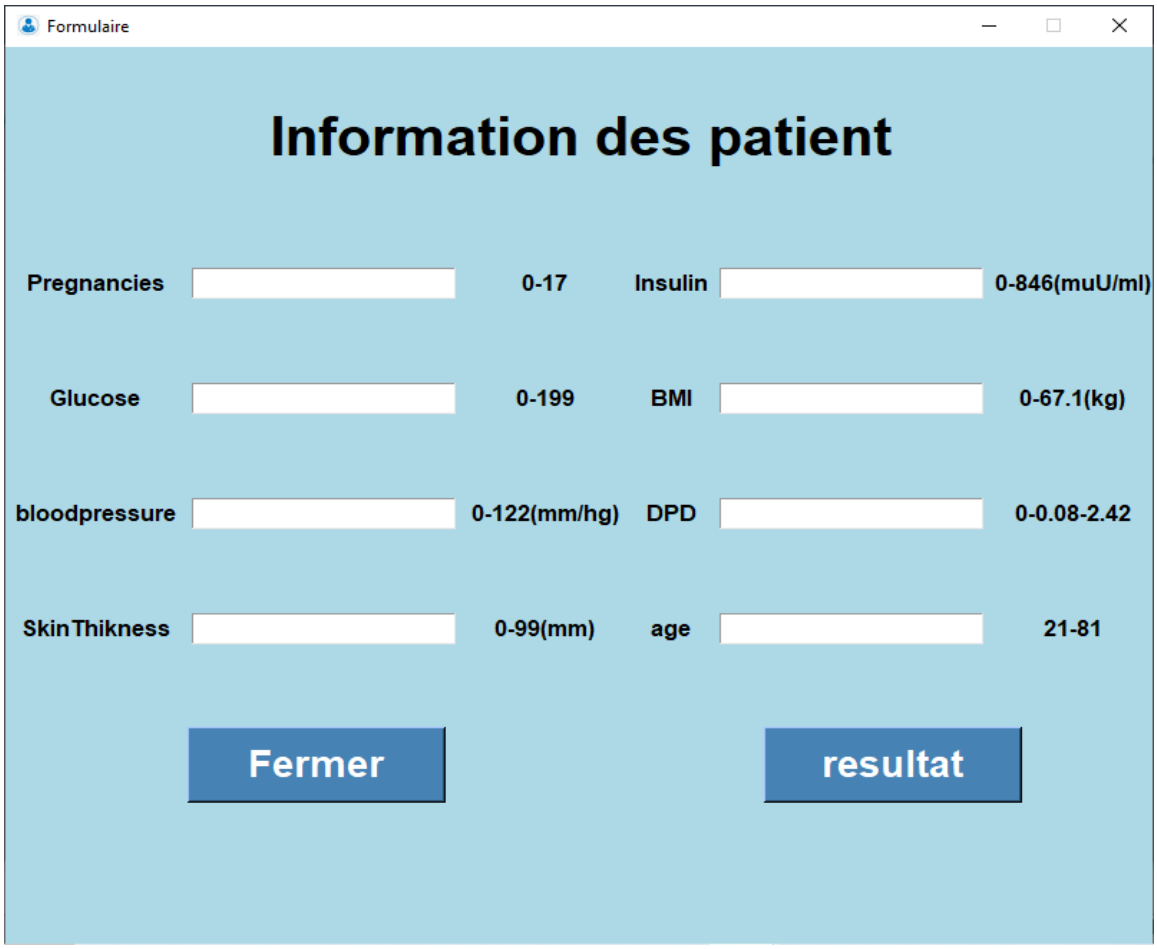


Figure 4-3:Formulaire du diabète

Après avoir saisi toutes les informations du patient, l'utilisateur peut cliquer sur le bouton « **résultat** » pour voir le résultat de diagnostic de ce patient en se basant sur les valeurs de ses symptômes saisies par l'utilisateur.

Le résultat s'affiche sous forme d'un **messagebox** comme l'indique la **figure 4.4**.

Si le patient est diabétique, un message de confirmation va afficher (**Figure 4-4**)

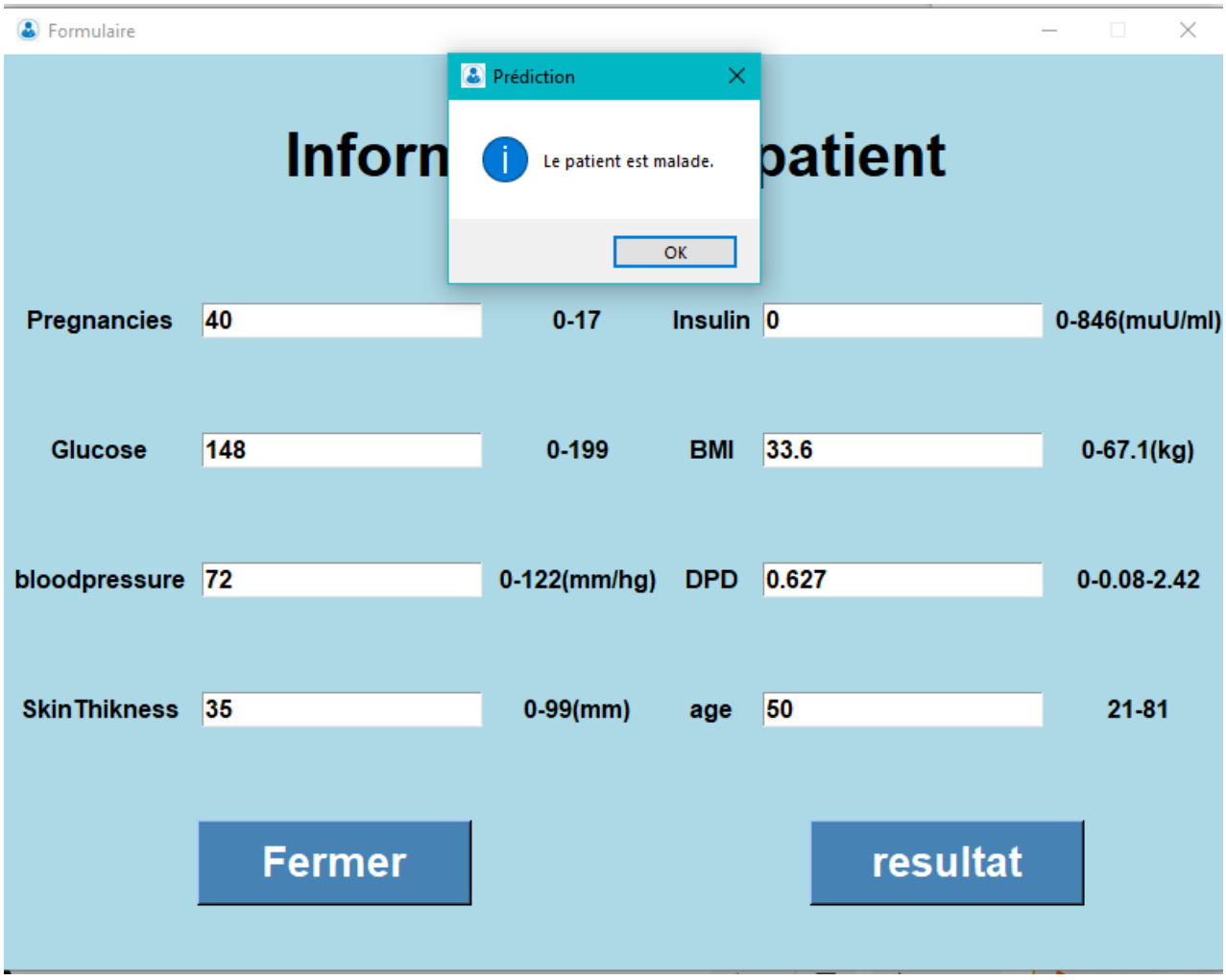


Figure 4-4:Message de confirmation de la maladie

Si le patient n'a pas une maladie de diabète, un message de négation va être affiché (**Figure 4.5**).

Figure 4-5: Message de négation de la maladie

4.3.3 Interface Maladie cardiaque : Message de négation de la maladie

En cliquant sur le bouton " **Maladie cardiaque** ", la **Figure 4.6** s'affiche.

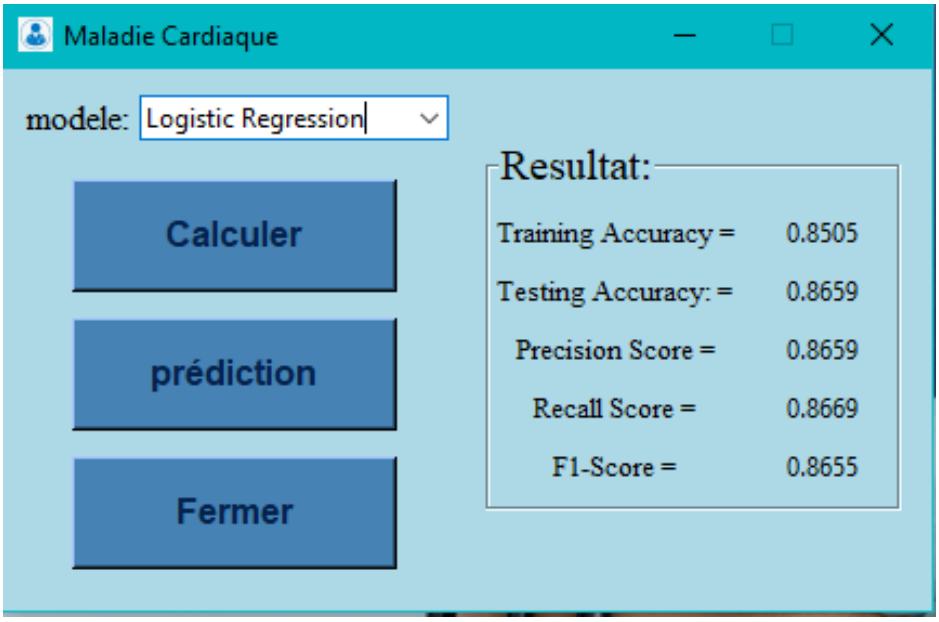


Figure 4-6: Précision d'apprentissage du Maladie cardiaque

Le bouton « **Prédiction** » sert à afficher la fenêtre de la **Figure 4.7** :

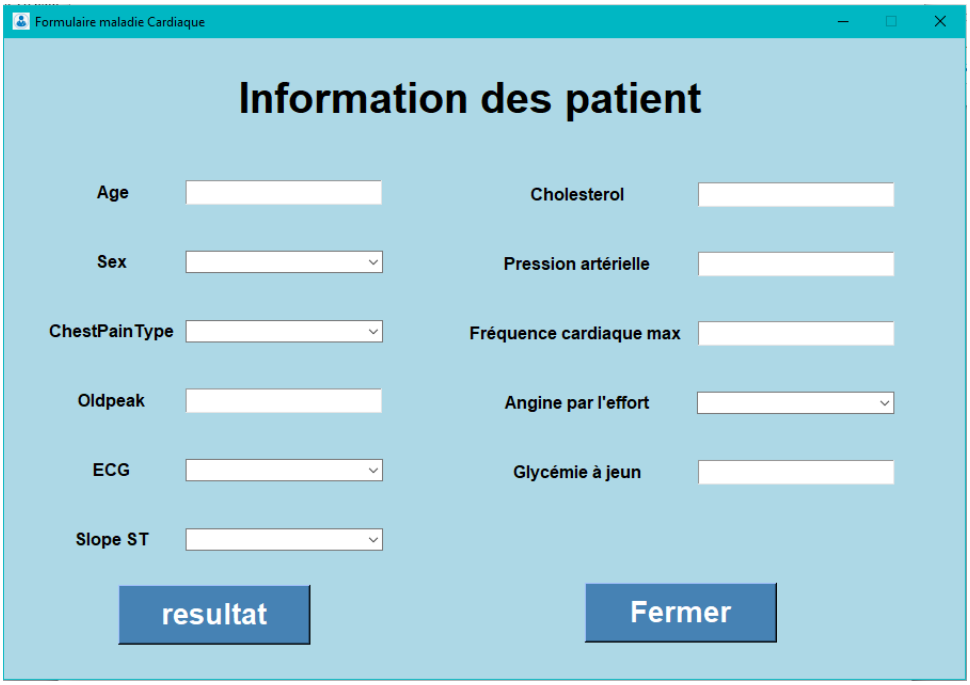


Figure 4-7:Formulair de la maladie cardiaque

Le bouton résultat lance la prédiction, ou bien le bouton Fermer pour vider le formulaire.

Si le patient a une maladie cardiovasculaire un message de confirmation va afficher

(Figure 4-8)

Figure 4-8 : Message de confirmation

Si le patient n’a pas une maladie cardiovasculaire un message de négation va afficher (**Figure 4-9**).

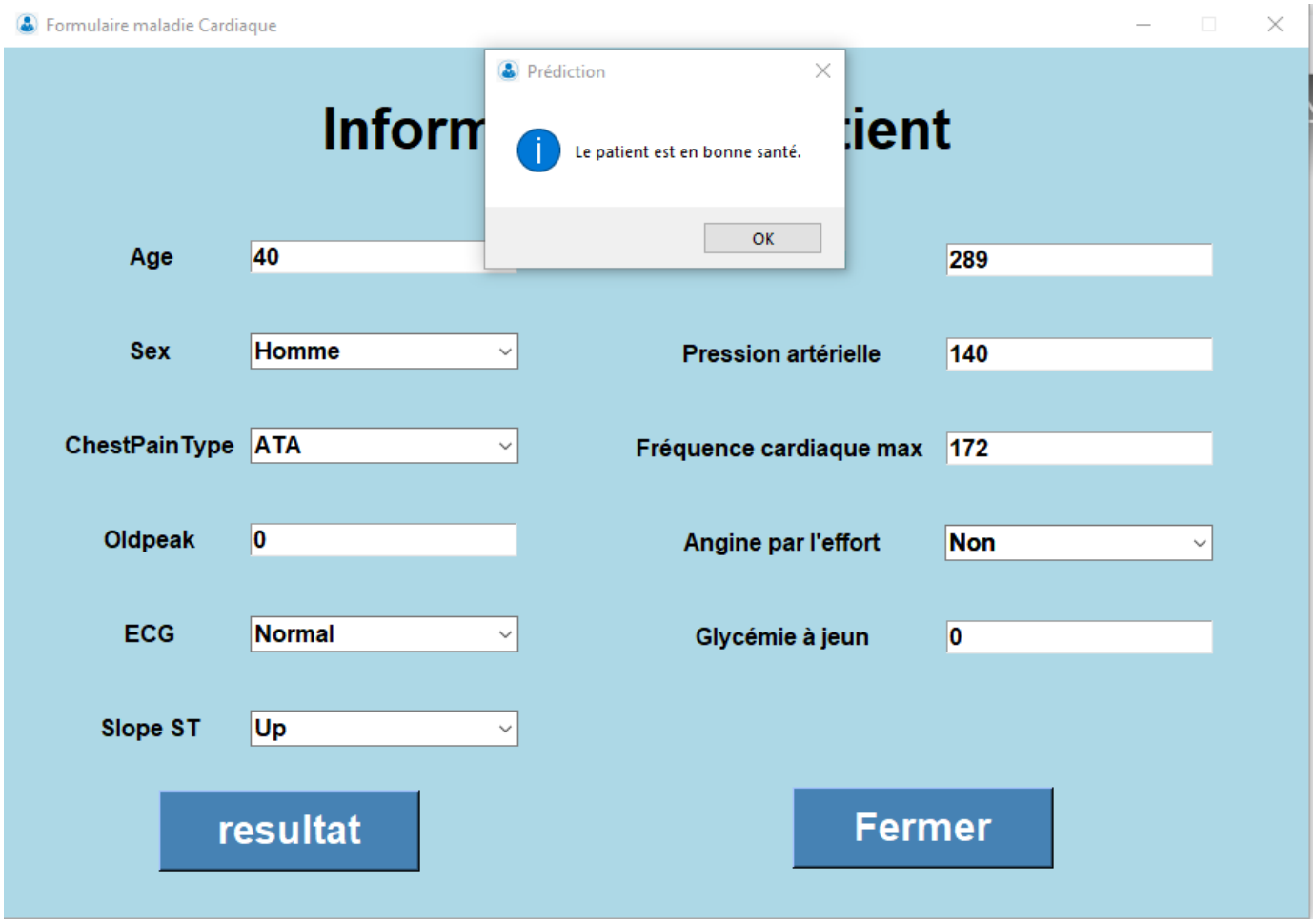


Figure 4-9: Message de négation de la maladie

Conclusion générale :

La prédiction des maladies à l'aide des techniques d'apprentissage automatique est un domaine prometteur de la médecine moderne. Les techniques d'apprentissage automatique peuvent être utilisées pour analyser les données des patients, apprendre les modèles de différentes maladies et prédire la probabilité qu'une maladie particulière se produise à l'avenir. Les applications de l'apprentissage automatique dans le domaine de la médecine représentent un grand pas vers l'amélioration des soins de santé. Elles permettent la détection précoce des maladies, ainsi qu'une amélioration du diagnostic et du traitement personnalisés. Les algorithmes d'apprentissage automatique peuvent aider à détecter les symptômes précoces des maladies et à prédire les risques de développer certaines pathologies. Cela permet aux médecins et aux professionnels de la santé d'agir plus rapidement pour diagnostiquer et traiter les patients, améliorant ainsi les résultats des traitements.

Dans cette thèse, nous nous intéressons à la conception et à la mise en œuvre d'une application qui compare les principales méthodes de classification supervisée sur les données médicales. La contribution de notre étude est de faciliter le travail du personnel médical, en particulier des médecins, en simplifiant la tâche de diagnostiquer ou de prédire les maladies. Nous espérons que les technologies futures continueront d'améliorer le processus médical en général et d'améliorer la qualité de vie des patients.

Ce projet a ouvert de nombreuses perspectives de travaux futurs, notamment l'amélioration de la qualité des soins de santé et la réduction des taux d'erreurs médicales, l'analyse et l'interprétation précise d'images médicales, ainsi que l'utilisation de l'apprentissage automatique pour améliorer la gestion des hôpitaux et des centres médicaux. Les applications de l'apprentissage automatique en médecine représentent un énorme changement dans l'amélioration des soins de santé et le développement du domaine de la médecine en général. Cependant, il convient de noter que la prédiction de la maladie à l'aide de techniques d'apprentissage automatique ne remplace pas un diagnostic médical formel. Au lieu de cela, ces techniques peuvent être utilisées comme outil complémentaire pour améliorer la capacité des médecins à diagnostiquer la maladie et à fournir de meilleurs soins de santé aux patients.

Références

[1] Conseil national de l‘Ordre des médecins, Conseil national de l‘Ordre des médecins: https://www.conseil-national.medecin.fr/.

[2] Livre de médecins et les pation, conseil-national.medcine.

[3] E. Topol, High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine,, 2019.

[4] Pensée Artificielle Machine Learning pour débutant : Introduction au Machine Learning.[en ligne]., Artificielle Machine Learning , 2020.

[5] Stat Soft [Stat Soft 06], Scikit-learn [Pedregosa 11] et le livre [Kelleher 15]., le livre [Kelleher 15]..

[6] java T point Support Vector Machine Algorithme, Récupéré sur java T point Support Vector Machine Algorithme.[en ligne].Disponible sur:: https ://www.javatpoint.com/machine-learning-support-vector-machine-algorithm.

[7] Gandhi, R.to wards data science. .Support Vector machine, Récupéré sur Gandhi, R.to wards data science. (2018).Support Vector Machine | Introduction to Machine Learning Algorithms.[en ligne].Disponible sur :: https ://towardsdatascience.com/support-vector-machine-introduction-to-machinelearning-algorithms-934a44, 2018.

[8] Exemple sur KNN :, Exemple sur KNN : : https://pixees.fr/informatiquelycee/n_site/nsi_prem_knn.html, 2018.

[9] Jason, 2013.

[10] De Wolf, L V Madden & P E Lipps., De Wolf, L V Madden & P E Lipps. Risk assessment models for, 2003.

[11] D. Hosmer Jr, Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression. John Wiley & Sons, 2013.

[12] liver de médecins et patients dans le monde des data des algorithmes et de l'intelligence artificielle, Récupéré sur Livre de médecins et patients dans le monde des data des algorithmes et de l'intelligence artificielle: https://www.conseil-national.medecin.fr

[13] «Appositeness of Optimized and Reliable Machine Learning for Healthcare».

[14] T. Quarter, «EHR,» *Electronic Health Records (EHR). (s.d.). American Journal of Health Sciences –Third Quarter 2012.*, 2012.

[15] T. Quater, « Electronic Health Records (EHR),» *Health Science,* 2012.

[16] «Appositeness of Optimized and Reliable Machine Learning,» 22 march 2022. [En ligne]. Available: https://doi.org/10.1007/s11831-022-09733-8.

[17] J. M. R. Imaging, M. A. Westwood et al. “Normalized left ventricular volumes and function in Thalassemia major patients with normal myocardial iron,, 2007.

[18] Sukanta Roy (le 19 Avril 2020). Accelerate Your Exploratory Data Analysis With Pandas-Profiling., [https://towardsdatascience.com/accelerate-your-explorato ry-data-analysis-with-pandas-profiling-4eca0cb770d1], (Consulté en 2021), 2021.

[19] Python - How and where to apply Feature Scaling., Python — How and where to apply Feature S[https://www.ge eksforgeeks.org/python-how-and-where-to-apply-feature-scaling/], (Consulté en 2021)., 2021.

[20] Using Pandas and Python to Explore Your Dataset, Using Pandas and Python to Explore Your Dataset . [https://realpython.com /pandas-python-explore-dataset/], (Consulté en 2021), 2021.

[21] "ML — Feature Scaling ", [https://www.geeksforgeeks.org/ml-feature-s caling-part-2/] , (Consulté en 2021)., 2021.

[22] «kaggele,» [En ligne]. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

[23] «kaggele,» [En ligne]. Available: https://www.kaggle.com/code/elouanguyon/pr-diction-maladies-cardiaques/notebook.

[24] P. S. A. K. Prof. Priya R. Patil, Marathwada Institute of "Technology., URL : https://www.ijariit.com/manuscripts/v3i2/V3I2-1197.pdf. , 2007.

[25] *(s.d.). Récupéré sur Conseil national de l'Ordre des médecins: https://www.conseil-national.medecin.fr/.*