République Algérienne Démocratique Et Populaire

Ministère de l'enseignement supérieur et de la recherche scientifique

Université du 8 mai 1945-Guelma-

Faculté des mathématiques, de l'informatique et des sciences de la matière

Département d'informatique

Thèse de Master

Spécialité : Informatique

Option:

Science et Technologie de l'Information et de la Communication

Thème

Un système intelligent pour améliorer la prédiction des maladies cardiovasculaires

Présentée par : Madjeda ZEMOULI

Membres du jury :

N°	Nom & prénom	Qualité
1	LOUAFI Wafa	Président
2	GUERROUI Nadia	Superviseur
3	BOUGHAREB Djalila	Examinateur

Soutenue le 26 juin 2023

Remerciements

Tout d'abord ,je rends grâce à Dieu le Tout-Puissant à qui j'exprime ma gratitude pour m'avoir donné,la force de relever la tête haute toute les épreuves et surtout pour réaliser ce travail .

Je tiens à remercier a mon encadreur Mme GUEROUI NADIA pour sa patience, sa-disponibilité et surtout ses judicieux conseils, qui ont contribué à alimenter ma réflexion.

Je tiens à remercier les membres du jury pour leur intérêt mené à mon projet en acceptant de passer en revue mon travail et l'enrichir de leurs propositions.

Je remercie le flambeau de ma vie et de mon soutien moral, ma chére mére. Et celui 'a qui je dois mon respect et mon amour est mon cher pére qui m'a donné courage et soutien psychologique..

Je voudrais remercier mon amie Wissal pour ses efforts et son soutien pour , et pour tous ces qui m'ont encouragé et soutenu pour surmonter les difficultés .

Grace a eux , j'ai pu avoir une base de travail solide sur laquelle j'ai pu m'appuyer pour réaliser ma démarche de recherche et d'analyse .

نظام ذكي لتحسين التنبؤ بأمراض القلب والأوعية الدموية

الملخص

إن الكشف المبكر عن أمراض القلب عامل حاسم للرعاية الصحية الناجحة. في السنوات الأخيرة ، شهد المجال الطبي ظهور طرق مختلفة للتنبؤ بأمراض القلب قبل حدوثها بناءً على التعلم الآلي والتعلم العميق. فعلا ، لا تزال أمراض القلب سبباً رئيسياً للوفيات في العالم. لذلك ، من الضروري تشخيصها وعلاجها في أسرع وقت ممكن. هدف هذا العمل هو تطوير نظام تنبؤ بأمراض القلب قادر على .كشف الظهور المبكر لمرض القلب الذي يمكن أن يكون قاتلا

المنهج المتبع في مشروع بحثنا هو الكشف عن أمراض القلب باستخدام تقنيات التعلم الآلي المشرفة. ولهذا الغرض، استكشفنا العديد من خوارزميات التعلم الآلي المشرفة لتحديد النموذج الأكثر فعالية للحصول على النتائج الأمثل. ركزنا دراستنا على تطبيق هذه النماذج على الكشف المبكر عن أمراض القلب، الأمر الذي يمكن أن يساعد على تحسين معدلات التشخيص المبكر ومنع المضاعفات المرتبطة بهذه الأمراض

تركّزعملنا على تحسين خوارزمية شجرة القرار للكشف عن أمراض القلب، يعتبر بالفعل أفضل خوارزمية في هذا المجال. بالإضافة إلى ذلك، قدّمنا تحسينات محددة على هذه الخوارزمية لزيادة كفاءتها. باستخدام نفس مجموعة البيانات المستخدمة في خوارزميات تعلّم الآلي المراقبة الأخرى، تمكّنا من إظهار أن خوارزميتنا المحسّنة أتت بنتائج أفضل. تؤكد هذه النتائج على فكرة أن نهجنا يوفّر حلًا .أكثر فعالية للكشف عن أمراض القلب مقارنة بالخوارزميات الحالية الأخرى

الكلمات الدالة : التنبؤ، تعلّم الآلة، أمراض القلب والأوعية الدموية، خوارزميات الذكاء الاصطناعي

Un système intelligent pour améliorer la prédiction des maladies cardiovasculaires

Résumé :

La détection précoce des maladies cardiaques est un facteur crucial pour une prise en charge réussie. Ces dernières années, le domaine médical a vu l'émergence de diverses méthodes basées sur l'apprentissage automatique et l'apprentissage profond pour prédire les maladies cardiaques avant qu'elles ne surviennent. En effet, les maladies cardiaques restent une cause majeure de décès dans le monde. Il est donc essentiel de les diagnostiquer et de les traiter le plus rapidement possible. L'objectif de ce travail est de développer un système de prédiction des maladies cardiaques capable de détecter l'apparition précoce d'une maladie cardiaque, qui peut s'avérer fatale.

La démarche de notre projet de recherche consiste à détecter les maladies cardiaques à l'aide de techniques d'apprentissage automatique supervisé. Pour ce faire, nous avons exploré plusieurs algorithmes d'apprentissage automatique supervisé afin d'identifier le modèle le plus efficace pour obtenir des résultats optimaux. Nous avons concentré notre étude sur l'application de ces modèles à la détection précoce des maladies cardiaques, ce qui pourrait contribuer à améliorer les taux de prédiagnostic tout en prévenant les complications associées à ces maladies.

Notre travail s'est concentré sur l'amélioration de l'algorithme de l'arbre de décision (DT) pour la détection des maladies cardiaques, qui était déjà considéré comme le meilleur algorithme dans ce domaine. En outre, nous avons proposé des améliorations spécifiques à cet algorithme afin d'accroître son efficacité. En utilisant le même ensemble de données que d'autres algorithmes d'apprentissage automatique supervisé, nous avons pu démontrer que notre algorithme amélioré produisait de meilleurs résultats. Ces résultats renforcent l'idée que notre approche offre une solution plus efficace pour la détection des maladies cardiaques que les autres algorithmes existants.

Mots clefs : Prédiction, apprentissage automatique, maladies cardiovasculaires, algorithmes d'intelligence artificielle.

A smart system to improve cardiovascular disease prediction

Abstract :

Early detection of heart disease is a crucial factor in its successful management. In recent years, the medical field has seen the emergence of various methods based on machine learning and deep learning to predict heart disease before it occurs. Indeed, heart disease remains a major cause of death worldwide. It is therefore essential to diagnose and treat them as quickly as possible. The aim of this work is to develop a heart disease prediction system capable of detecting the early onset of heart disease, which can prove fatal.

This research project's approach aims to detect heart disease using supervised machine learning techniques. To this end, we explored several supervised machine learning algorithms in order to identify the most efficient model for achieving the optimum results. Specifically, we focused on how to apply our proposed model approach to the early detection of heart disease, which may help to improve pre-diagnosis and reduce the risks of complications of heart disease.

Our work focused on improving the decision tree (DT) algorithm for heart disease detection, which was already considered the best algorithm in this field. In addition, we proposed specific improvements to this algorithm to increase its efficiency. Using the same dataset as other supervised machine learning algorithms, we were able to demonstrate that our improved algorithm produced better results. These results reinforce the idea that our approach offers a more effective solution for heart disease detection than other existing algorithms.

Keywords : Prediction, machine learning, cardiovascular disease, artificial intelligence algorithms..

Dédicace

"À tous ceux qui ont rendu mon parcours de projet de fin d'études (PFE) inoubliable "

Je n'oublierai jamais les liens indéfectibles tissés avec mes camarades de classe. Ensemble, nous avons partagé des moments d'apprentissage, de travail acharné et de rires. Leurs encouragements, leur esprit d'équipe et leur soutien mutuel ont été une source constante de motivation.

"Je dédie mon diplôme à l'âme de mon cher grand-père, que Dieu ait pitié de lui".

Enfin, je souhaite adresser une dédicace spéciale à ma famille zemouli et hachemi et à mes proches .

Leur amour inconditionnel, leur soutien indéfectible et leur compréhension ont été mes piliers tout au long de ce projet. Leur présence m'a permis de rester concentré et de garder le cap, même dans les moments les plus difficiles.

Table des matières

Liste des tableaux

Table des figures

Introduction générale

Contexte de la thèse

Les maladies cardiaques sont devenues l'une des maladies les plus préoccupantes au monde, avec des conséquences dramatiques pour la santé des personnes. Ces dernières années, elles sont devenues la première cause de mortalité dans le monde. Pour éviter que les patients ne subissent d'autres dommages, il est essentiel de diagnostiquer les maladies cardiaques avec précision et en temps utile. Récemment, des techniques médicales innovantes, telles que les techniques basées sur l'intelligence artificielle, ont été utilisées dans le domaine médical.

Problématique de la thèse

Dans ce domaine, l'apprentissage automatique tire parti d'algorithmes et de techniques utiles qui sont souvent utilisés pour diagnostiquer les maladies cardiovasculaires plus rapidement et avec plus de précision. Cependant, prédire les maladies cardiaques reste une tâche compliquée, voire impossible, pour les médecins et les professionnels de la santé.

Face à un volume important de données médicales, il est difficile pour les spécialistes de comprendre et de prédire les aspects complexes des maladies cardiaques.

Dans cette optique, le recours à des techniques avancées d'apprentissage automatique et d'intelligence artificielle offre de nouvelles possibilités pour mieux anticiper et prendre en charge ces maladies.

Contributions de la thèse

Cette thèse s'inscrit dans le domaine de la recherche médicale et de l'informatique de santé. Elle explore le potentiel des algorithmes d'apprentissage automatique tels que l'arbre de décision (DT), les K-voisins les plus proches *(KNN)*, la machine à vecteur de support *(SVM)* et la régression logistique *(LR)* pour améliorer la prédiction des maladies cardiovasculaires. Ces algorithmes sont largement utilisés dans le domaine de l'apprentissage automatique et ont montré des performances prometteuses dans diverses applications.

Ce travail apporte une contribution significative au domaine de la prédiction des maladies cardiovasculaires. Grâce à un ensemble de données spécifiques, nous avons pu démon-

trer que l'algorithme de l'arbre de décision atteint un taux de précision remarquable de 99,02 % dans la prédiction des maladies cardiovasculaires. Par ailleurs, cette étude évalue et analyse les performances d'algorithmes d'apprentissage automatique tels que l'arbre de décision, le K-voisin le plus proche, la machine à vecteur de support et la régression logistique.

En outre, cette thèse propose une approche pour calculer la probabilité de maladie cardiovasculaire en utilisant le modèle d'apprentissage par arbre de décision. Cette probabilité peut être utilisée comme référence pour estimer le risque qu'a une personne de développer une maladie cardiovasculaire.

Organisation de la thèse

Outre l'introduction générale qui précise le contexte de l'étude, la problématique de recherche, les contributions de la thèse ainsi que l'organisation du manuscrit,le manuscrit est divisé en trois chapitres comme suit :

Chapitre 1 : **État de l'art.** Il offre au lecteur un aperçu rapide et complet du domaine de la recherche. Il est consacré à tout ce qui concerne les maladies cardiovasculaires, y compris les différents types de maladies, leurs causes, leurs signes et leurs caractéristiques. Les concepts relatifs à l'apprentissage automatique et à la recherche sur la prédiction des maladies cardiaques sont passés en revue, de même que la littérature sur la prédiction des maladies cardiovasculaires.

Chapitre 2 : **La conception d'un système intelligent de prédiction des maladies cardiovasculaire**. Une proposition de système de prédiction des maladies cardiovasculaires est présentée. Pour ce faire, un algorithme d'arbre de décision est utilisé pour la classification.

Chapitre 3 : **Implémentation et résultats expérimentaux.** Les différents outils et langages de développement permettant l'utilisation d'algorithmes ont été présentés, ainsi que la solution que nous avons proposée pour l'optimisation **DT**, où nous avons comparé les résultats de prédiction de certains des algorithmes étudiés avec ceux de l'algorithme amélioré. Nous avons démontré comment l'approche proposée apporte une valeur ajoutée au domaine et surpasse les approches existantes.

Enfin, la conclusion générale rappelle l'intérêt de la démarche, souligne ses limites et propose quelques perspectives de recherche future basées sur les principaux résultats dégagés de cette thèse.

Chapitre 1

État de l'art

1.1 Introduction

Les maladies cardiovasculaire regroupent les pathologies qui touchent le cœur et l'ensemble des vaisseaux sanguins, comme l'athérosclérose, les troubles du rythme cardiaque, l'hypertension artérielle, l'infarctus du myocarde, l'insuffisance cardiaque ou encore les accidents vasculaires cérébraux. Dans ce chapitre, nous en apprenons davantage sur les maladies cardiaques en identifiant leurs types, causes et symptômes. Les maladies cardiaques.

1.2 Concepts généraux et définitions

1.2.1 Les maladies cardiaques

Les maladies cardiovasculaires sont pour l'essentiel secondaires à l'athérosclérose dont le développement est très dépendant de notre mode de vie occidental, Ceux-ci finissent par gêner, voire empêcher la circulation du sang qui alimente le cœur, le cerveau ou les jambes, provoquant des angines de poitrine, des infarctus, des AVC, et des artérites [Baudet et al., 2012].

1.2.2 Types des maladies cardiaques

Les maladies cardiovasculaires comprennent plusieurs types de troubles de l'appareil circulatoire, à savoir les maladies congénitales, notamment [W.h.organisation, 2022] :

- Les cardiopathies coronariennes : elles affectent les vaisseaux sanguins qui alimentent le muscle cardio-vasculaire.
- Les maladies cérébro-vasculaires : affectant les vaisseaux sanguins alimentés par le cerveau.
- Les artériopathies périphériques : affectant les vaisseaux sanguins qui fournissent les bras et les pattes.
- Les cardiopathies rhumatismales : affectant les valvules musculaires et cardiaques et résultant de rhumatismes articulaires aigus, provoqués par les bactéries streptocoques.

- Les malformations cardiaques congénitales : malformations de la structure cardiaque déjà existantes à la naissance.
- Les thromboses veineuses profondes et les embolies pulmonaires : blocage des veines dans les jambes par un caillot de sang, qui peut se détacher et migrer vers le cœur ou les poumons.

1.2.3 Les origines des maladies cardiaques

Les maladies cardiovasculaires **(MCV)** résultent généralement d'une accumulation de dépôts lipidiques sur les parois des artères coronaires, entraînant une diminution du flux sanguin vers le cœur. On peut citer parmi les différents facteurs de risque de problèmes cardiaques [Baudet et al., 2012] :

— *Le tabac* : Le fait d'arrêter de fumer vous permettra non seulement de réduire votre risque de problèmes cardiaques, mais vous procurera également de nombreux autres bénéfices.

— *Le taux de cholestérol élevé* : une quantité trop élevée de cholestérol dans le sang constitue un des principaux facteurs de risque de maladie cardiovasculaire .

— *La pression artérielle élevée* : la pression artérielle mesure la force avec laquelle le sang est expulsé depuis le cœur vers le reste du corps, à travers les artères.

— *Le surpoids ou l'obésité* : se situer au-dessus du poids recommandé peut entraîner l'apparition d'autres facteurs de risque de maladie du cœur.

— *Le diabète* : Si une personne est atteinte de diabète, elle doit commencer à obtenir l'aide dont elle a besoin pour traiter cette maladie et à prendre les précautions nécessaires pour agir contre les autres causes possibles des maladies cardio-vasculaires.

— Une alimentation non équilibrée contribue a hauteur de 33% au risque d'AVC .

— La fibrillation atriale, qui est le premier facteur de risque d'origine cardiaque, avec un risque multiplié par quatre (04).

— La consommation d'alcool

1.2.4 Les recommandations pour les maladies cardio-vasculaires

Il existe de nombreux moyens de prévention contre les maladies cardiovasculaires. On peut citer, entre autres [Rivière, 2019], les suivantes :

— *Arrêter de fumer :* La recherche montre que le tabagisme augmente le risque de maladie cardiaque et pulmonaire, de divers cancers, d'affaissement de la peau et d'accélération du vieillissement, d'accident vasculaire cérébral et de démence précoce.

— *Adopter une alimentation saine et équilibrée :* Le mauvais cholestérol **(LDL)** est le facteur numéro un que nous devons maintenir dans la plage normale, et nous devons également éviter les triglycérides élevés,Comme les médecins conseillent de manger des légumes, des fruits, du poisson, des produits alimentaires à base de soja, des produits laitiers, des produits alimentaires faibles en gras riches en fibres, et les médecins avertissent de la consommation fréquente de viande, de produits de boulangerie, de produits laitiers et de restauration rapide riches en graisses saturées, vous devez également limiter la consommation de glucides présents dans le pain

blanc, les pommes de terre, le chocolat, la carotte, la pastèque, le riz blanc, les bonbons, les aliments frits et les aliments riches en sel.

— *lcool :* L'alcool est un sucre qui peut pénétrer directement dans le sang, fait augmenter le taux de triglycérides dans le sang et affecte négativement l'hypertension. Il est préférable de s'abstenir de consommer de l'alcool et de recourir à des moyens plus sains pour augmenter le taux de bon cholestérol .

— Surveiller son poids.

— Lutter contre la sédentarité.

— Surveiller les taux de cholestérol ainsi que les triglycérides.

— *vérifier la tension artérielle régulièrement :* Il n'y a pas de médicament pour se débarrasser du sentiment de stress et de stress une fois pour toutes, il existe des techniques de relaxation telles que le yoga et la méditation, en plus d'un bon sommeil et de repos, créant un environnement favorable qui comprend des amis et des proches pour partager avec eux nos problèmes lorsque nous sommes bouleversés.

— Prendre le contrôle de son diabète.

— La prévention, un enjeu majeur.

1.2.5 Les signes de maladie cardiaque

Plusieurs symptômes indiquent une plus grande chance de contracter une maladie cardiaque, à savoir les suivants :

- Certains types de douleur ;
- Fatigue ;
- Palpitations ;
- Évanouissement ;
- Gonflement des jambes, des chevilles et des pieds ;
- Essoufflement.

Ces signes ne sont pas suffisants pour confirmer la maladie, c'est pourquoi nous devons procéder à de nombreux tests et examens médicaux, qui nous fournissent de nombreux attributs essentiels contenant davantage d'informations qui nous permettent de détecter la maladie.

1.2.6 Les attributs de la maladie

Afin de diagnostiquer les maladies cardiovasculaires, il est nécessaire de collecter autant d'informations que disponibles dans la base de données de Heart Disease Dataset [Smith, 2019], qui contient 76 attributs et 1025 instant , y compris l'attribut prévu , mais toutes les expériences publiées font référence a l'utilisation d'un sous-ensemble de 14 d'entre eux, et d'identifier les symptômes pour assurer un diagnostic exact. La tabe 1.1 présente les attributs choisis et les informations correspondantes (*le cas échéant*).

Ces paramètres ont été retenus dans le cadre d'autres recherches et par des spécialistes de la santé, estimant que ces attributs représentent les meilleurs indicateurs pour les maladies cardio-vasculaires.

Attributs	**Indications**
Age	en anneés
Sex	sexe (1 = masculin ; 0 = féminin).
ChestPainType	chest pain type (0= TA : TypicalAngina, 1= ATA : AtypicalAngina, 2 = NAP :non-Anginal Pain,3 = ASY : Asymptomatic)
Resting blood pressure	in mm Hg on admission to the hospital
Serum cholestoral	mg/dl
Fasting blood sugar	fbs > 120 mg/dl (1 = true ; 0 = false)
Resting electrocardiographic results	values repos (**0**= normal ; **1**= Abnormal ; **2**= Hypertrophy)
Maximum heart rate achieved	
Exercise induced angina	1 = yes ; 0 = no
Oldpeak	= *ST depression induced by exercise relative to rest*
the slope of the peak exercise ST segment	
number of major vessels (0-3) colored by flourosopy	
thal	: **0** = *normal* ; **1** = *fixed defect* ; **2** = *reversable defect*

TABLE 1.1 – Représentation des attributs de tous les éléments du jeu de données .

1.3 L'apprentissage automatique (principe et application)

1.3.1 Principe

Les techniques d'apprentissage automatique font référence aux algorithmes et aux modèles statistiques qui permettent aux machines d'apprendre à partir de données sans être explicitement programmées. Ces techniques reposent sur les principes de l'inférence statistique, de la reconnaissance des formes et de l'optimisation. L'objectif est de construire des modèles prédictifs capables d'identifier des modèles et des relations dans les données, et d'utiliser ces connaissances pour prendre des décisions ou faire des prédictions sur de nouvelles données, Voici le shema suivante : Figure1.1 [Prasad Lokulwar, 2022].

Il existe quatre principaux modèles de techniques d'apprentissage automatique : l'apprentissage supervisé, l'apprentissage non supervisé, l'apprentissage semi-supervisé et l'apprentissage par renforcement [Prasad Lokulwar, 2022]. L'apprentissage supervisé implique l'utilisation de données étiquetées pour entraîner un modèle à faire des prédictions sur de nouvelles données encore inconnues. L'apprentissage non supervisé, quant à lui, implique l'utilisation de données non classées pour découvrir des modèles et des relations au sein des données. L'apprentissage semi-supervisé combine des données étiquetées et non étiquetées pour former un modèle, et l'apprentissage par renforcement implique l'utilisation de récompenses et de sanctions pour former un modèle à prendre des décisions [Zhang, 2010].

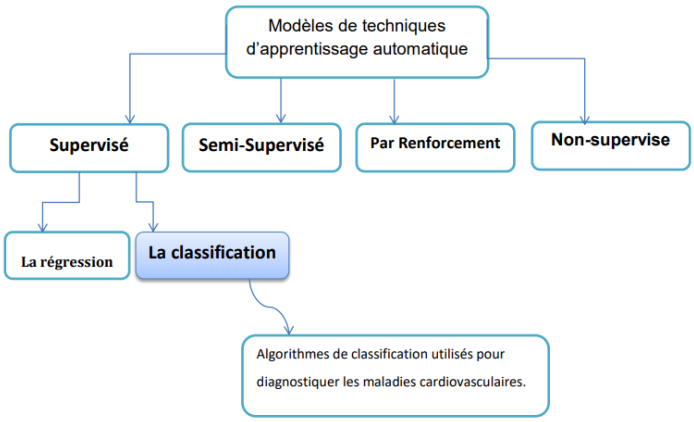


FIGURE 1.1 – Les modèles de techniques d'apprentissage automatique.

1.3.2 Application

Les techniques d'apprentissage automatique ont un large éventail d'applications dans divers domaines tels que la vision par ordinateur, le traitement du langage naturel, la reconnaissance vocale, la robotique et les soins de santé. Dans le domaine de la vision artificielle, les algorithmes d'apprentissage automatique sont utilisés pour le traitement d'images et de vidéos, la détection et la reconnaissance d'objets [Jain, 2023]. Dans le traitement du langage naturel, l'apprentissage automatique est utilisé pour l'analyse des sentiments, la classification des textes et la traduction des langues. Dans le domaine de la reconnaissance vocale, l'apprentissage automatique est utilisé pour reconnaître et transcrire la parole en texte. En robotique, les algorithmes d'apprentissage automatique sont utilisés pour la planification de trajectoires, la reconnaissance d'objets et le contrôle. Dans le domaine de la santé, les techniques d'apprentissage automatique sont utilisées pour le diagnostic des maladies, la découverte de médicaments et la médecine personnalisée. Dans l'ensemble, les techniques d'apprentissage automatique offrent des outils puissants pour l'analyse des données et la prise de décision dans un grand nombre de domaines. [Zhang, 2010] et voila quelques algorithmes de classification utilisés pour diagnostiquer les maladies cardiovasculaires :

1.3.3 Support Vector Machine SVM

Le **SVM** est un modèle d'apprentissage automatique utilisé à la fois pour la classification et la régression. C'est un algorithme supervisé qui vise à trouver un hyperplan optimal pour séparer les exemples de différentes classes dans un espace de dimension supérieure.

L'objectif est de trouver l'hyperplan qui maximise la marge entre les exemples de différentes classes. La marge est la distance entre l'hyperplan et les exemples les plus proches de chaque classe, appelés vecteurs de support. Ces vecteurs de support sont les points de données les plus difficiles à classifier et jouent un rôle crucial dans la définition de l'hyperplan [Mahesh, 2020]*(voir la figure1.2)*.

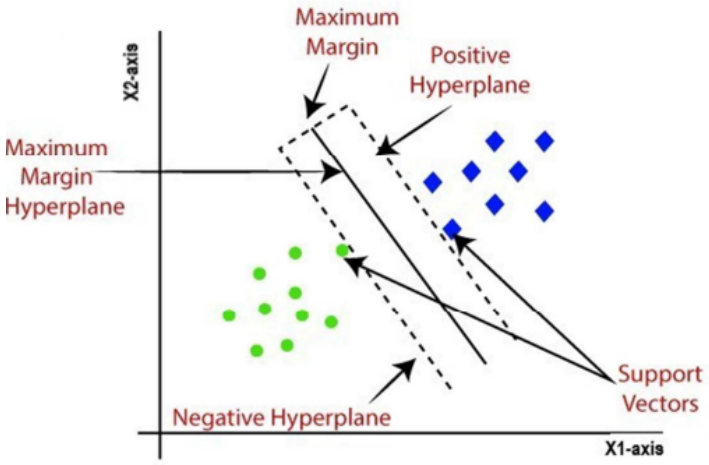


FIGURE 1.2 – Support Vector Machine

1.3.4 Arbre de décision (DT)

L'arbre de décision est un modèle d'apprentissage automatique couramment utilisé en apprentissage supervisé *(voir la figure 1.3)*. Il est particulièrement utile pour la classification et la régression [Mahesh, 2020].

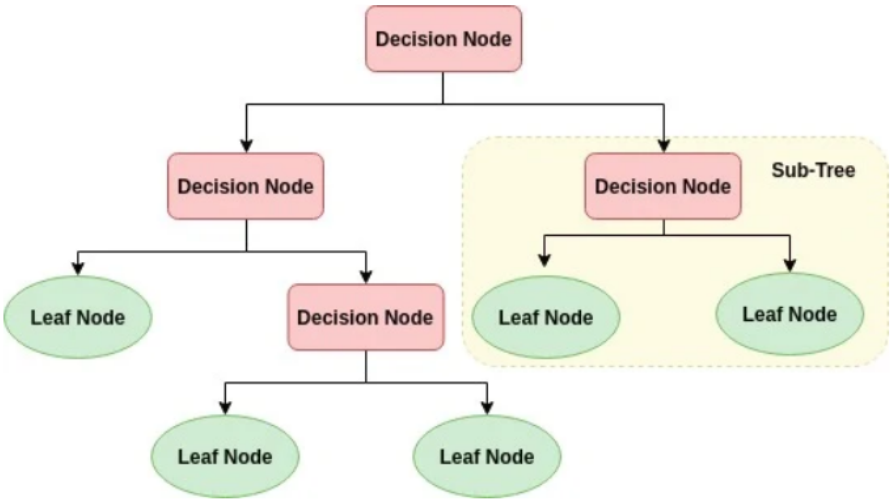


FIGURE 1.3 – L'arbre de décision (*Tree decision*).

Objectif : Créer un modèle qui prédit la valeur d'une variable cible en apprenant de simples règles de décision dérivées des données d'entraînement. L'arbre de décision commence par un nœud racine qui représente l'ensemble des données d'entraînement. À partir de ce nœud, l'algorithme recherche les attributs qui ont la plus grande capacité à diviser les données en groupes homogènes en termes de variable cible. Cela signifie que l'algorithme cherche à diviser les données en groupes qui ont la même valeur de variable cible.

Avantages : Les avantages de l'utilisation d'un arbre de décision comprennent sa simplicité, sa facilité d'interprétation et sa capacité à gérer des données manquantes. Ce-

pendant, les arbres de décision peuvent également être sensibles au sur-apprentissage et nécessitent souvent une certaine forme de régularisation pour améliorer leur performance sur des données de test.

1.3.5 Apprentissage avec les arbres de décision

Considérons d'abord le problème de la classification. Chaque élément x de la base de données est représenté par un vecteur multidimensionnel correspondant l'ensemble de variables descriptives d'un point. Chaque nœud interne de l'arbre correspond à un test effectué sur l'une des variables :

— *Variable catégorielle :* Crée une branche (descendante) par valeur d'attribut.

— *Variable numérique :* Test par intervalles de valeurs.

L'utilisation des feuilles de l'arbre permet de spécifier les classes et d'encoder la règle de décision. Une fois que l'arbre est construit, la classification d'un nouvel individu se fait en suivant un chemin descendant de la racine vers l'une des feuilles. À chaque niveau de la descente, un nœud intermédiaire est traversé, où une variable est testée pour déterminer le chemin à prendre pour poursuivre la descente [Crucianu and Philippe, 2010].

1.3.6 Les K plus proches voisins (KNN)

est un algorithme d'apprentissage automatique utilisé principalement pour la classification, bien qu'il puisse également être utilisé pour la régression. Il appartient à la famille des algorithmes d'apprentissage supervisé *(Voir la Figure 1.4)* :

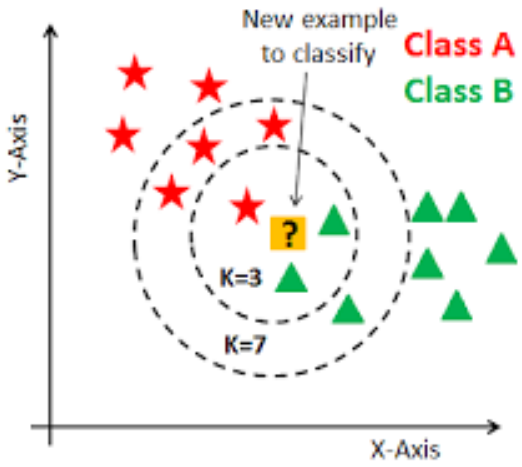


FIGURE 1.4 – Utilisation de la distance dans l'algorithme KNN.

Objectif est de prédire l'étiquette de classe (dans le cas de la classification) ou la valeur numérique (dans le cas de la régression) d'un nouvel exemple non étiqueté en se basant sur les étiquettes des exemples étiquetés les plus proches dans l'espace des caractéristiques.

Cependant, il est important de noter que KNN peut avoir des limitations en termes de temps de calcul, en particulier avec de grands ensembles de données, car il nécessite de calculer la distance entre le nouvel exemple et tous les exemples d'entraînement. De plus,

il peut être sensible à la curse of dimensionalité (*malédiction de la dimension*) lorsque l'espace des caractéristiques est de grande dimension. [Kumar et al., 2021]

1.3.7 Naïve Bayes (NB)

La base du classificateur bayes naïf est bayes théorème. Une hypothèse est générée pour l'ensemble donné de classes.

— le Naive Bayes calcule la probabilité a posteriori de chaque classe pour une nouvelle observation non étiquetée en utilisant le théorème de Bayes, et choisit la classe avec la plus grande probabilité a posteriori comme prédiction.

— L'un des avantages majeurs du Naive Bayes est sa simplicité et sa rapidité de calcul. De plus, il peut fonctionner efficacement même avec des ensembles de données de grande dimension.

Dans l'hypothèse de l'indépendance de l'algorithme bayésien naïf est faite. Sur la base de la valeur cible, les valeurs de l'attribut sont choisies et elles sont indépendantes d'une seule [Ramalingam et al., 2018]. Le théorème de Bayes fournit un moyen de calculer la probabilité a posteriori $P(c|x)$ à partir de $P(c)$, $P(x)$ et $P(x|c)$ avec :
$P(c|x) = \frac{P(c|x).P(x)}{P(x))}$ et $P(c|x)) = P(x_1|c).P(x_2|c)....P(x_n|c).P(c)$

Où : $P(c|x)$: la probabilité de l'événement c sachant que l'événement x s'est produit. $P(c)$: est la probabilité de l'événement c indépendamment de x. $P(x)$: est la probabilité de l'événement x.

1.3.8 La régression logistique (LR)

La régression logistique est un algorithme d'apprentissage automatique utilisé pour la classification binaire, c'est-à-dire la prédiction de deux classes distinctes *(voir la figure 1.5)*.
Dans ce modèle, les caractéristiques de l'exemple sont pondérées par des coefficients (poids) et sommées pour obtenir une valeur continue appelée score. Ensuite, la fonction logistique est appliquée à ce score pour obtenir une valeur de probabilité comprise entre **0** et **1** [Kumar et al., 2021].

La régression logistique présente plusieurs avantages, tels que sa simplicité, son interprétabilité et sa rapidité de calcul. Elle est également robuste face à des ensembles de données de grande taille et peut être étendue à des problèmes de classification multi-classe.

1.4 L'apprentissage automatique au service des maladies cardiaques

L'apprentissage automatique est une branche de l'intelligence artificielle qui se concentre sur le développement d'algorithmes capables d'apprendre à partir de données et de faire des prédictions. Dans le contexte des maladies cardiaques, l'apprentissage automatique consiste à analyser de vastes ensembles de données sur les patients afin d'identifier des schémas et des relations entre les variables. Ces algorithmes utilisent des modèles statistiques

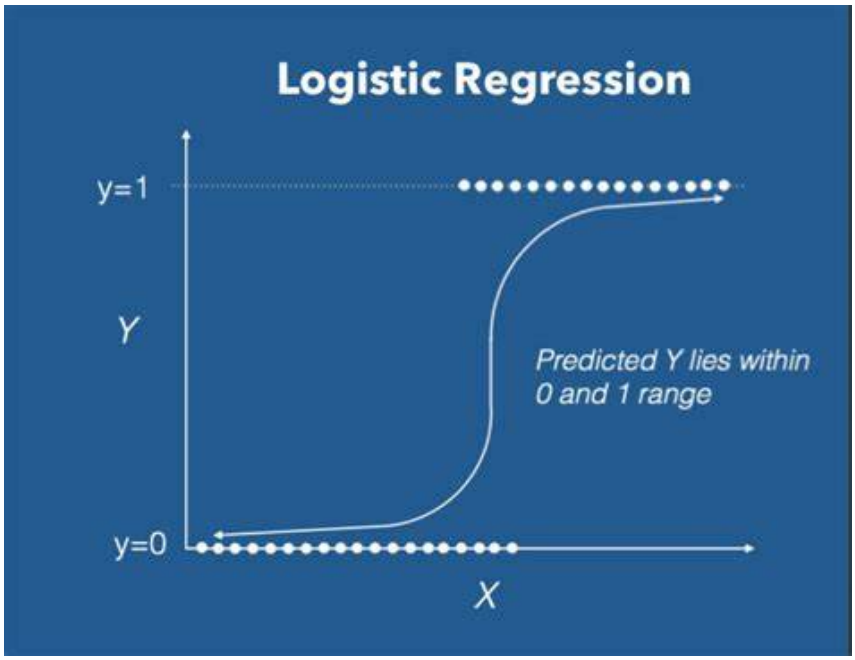


FIGURE 1.5 – La régression logistique LR.

pour identifier les facteurs de risque, prédire les résultats et faire des recommandations de traitement [Fayez, 2023].

Les modèles d'apprentissage automatique sont formés à partir des données recueillies auprès des patients, puis soumis à des tests sur de nouvelles données afin d'en mesurer la précision et l'efficacité. L'objectif de l'apprentissage automatique dans le domaine des maladies cardiaques est d'améliorer la précision du diagnostic, de prédire la probabilité d'événements futurs et, en fin de compte, d'améliorer les résultats pour les patients [Bhatt et al., 2023a].

1.4.1 Mesure de performance

Les mesures de performance sont des indicateurs de la qualité de la correspondance entre Valeurs prédites et valeurs obtenues à partir du modèle obtenu via la matrice de confusion [Weng, 2020]. Comme le montre la table 1.2 suivante :

		Predicted		
		True	False	
Actual	True	True positive (**TP**)	False negative (**FN**) Type II Error	$Recall = Sensntivity = \frac{TP}{TP+FN}$
	False	False positive (**FP**) Type I Error	True negative (**TN**)	$Specificity = \frac{TN}{TN+FP}$
		$Precision = \frac{TP}{TP+FP}$		$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$; $F1 = \frac{2*Precision*Recall}{Precision+Recall}$

TABLE 1.2 – Métriques d'évaluation de performance

1.5 Les travaux connexes

Dans [Son et al., 2010],un algorithme d'apprentissage automatique Support Vector Machine (SVM) a été utilisé pour la prédiction de l'insuffisance cardiaque. Pour ce faire, le travail a porté sur un ensemble de données de 11 variables provenant d'un certain nombre de patients souffrant d'insuffisance cardiaque. La robustesse du modèle a été vérifiée à l'aide de la méthode LOOCV (Leave One Out Cross Validation).

Les auteurs ont obtenu une précision maximale de **77,63** avec ce modèle. Les recherches menées dans le cadre de l'étude sur les maladies cardiaques [Polat et al., 2007] ont utilisé le "k-nearest neighbor" (k-nn) pour diagnostiquer les maladies cardiaques. Le diagnostic des maladies cardiaques a été effectué à l'aide d'un mécanisme d'apprentissage automatique. Pour cela, avant le classificateur principal, une étape de prétraitement a été réalisée à l'aide de la méthode des voisins les plus proches (K-NN). Il était évident que les algorithmes de classification étaient devenus populaires dans le diagnostic des maladies cardiaques. La précision du système offert a été de 87%.

Dans [Smith, 2019] une base de données *Heart Disease Prediction* pour prédire les possibilités d'apparition de maladies cardiaques chez les patients, l'algorithme DT s'est avéré le mieux adapté à cette classification (pour une précision de 99%). Ce problème apparaît notamment lorsque le nombre de variables d'entrée est relativement plus important que le nombre d'observations.

Dans [Deepika and Sasikala, 2020] la précision pour leur modèle est de 83,6%. Il s'agit de l'une des précisions élevées produites par le modèle de l'arbre de décision. Le temps pris pour cette prédiction en utilisant le modèle d'arbre de décision est de *0.02 secondes*.

Les travaux de [Rubini et al., 2021] ont présenté une analyse comparative des techniques d'apprentissage automatique telles que le classificateur Random Forest **(RFC)**, la régression logistique **(LR)**, la machine à vecteurs de support (SVM) et Naïve Bayes (NB) dans la classification des maladies cardiovasculaires. Leur analyse comparative a révélé que le RFC et le LR ont obtenu les meilleures précisions, soit 84,81 % et 83,82 %.

La table 1.3 illustre certains travaux récents dans ce domaine.

1.6 Conclusion

Ce chapitre est consacré à tout ce qui concerne les maladies cardiovasculaires, notamment les différents types de maladies, leurs causes, leurs signes et leurs caractéristiques. Nous avons passé en revue les concepts de l'apprentissage automatique et les recherches réalisées sur la prédiction des maladies cardiaques, et nous avons aussi fait le point sur les travaux réalisés dans la littérature sur la prédiction des maladies cardio-vasculaires. Dans le prochain chapitre, nous présenterons en détail notre approche suggérée pour prédire les maladies cardiaques à l'aide d'algorithmes d'apprentissage automatique.

Année	Auteur	Techniques	Accuracy	Observation
2010	[Son et al., 2010]	SVM	Svm = 77.63%	machine learning algorithm Support Vector Machine (SVM) for heart failure prediction
2016	[Saqlain et al., 2016]	DT, LR, RF, ANN , NB , SVM	NB=86.6%	Identification of Heart Failure by Using Unstructured Data of Cardiac Patientss
2018	[Polat et al., 2007]	Adaboost, DT, LR, RF, SVM	LR=87.1%	Application de l'apprentissage automatique à la prédiction des maladies
2019	[Ali et al., 2019]	SVM linear , SVM RBF	SVM =92,22%	An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure
2020	[Deepika and Sasikala, 2020]	DT	DT= 83,6%	Enhanced Model for Prediction and Classification of Cardiovascular Disease using Decision Tree with Particle Swarm Optimization
2021	[Rubini et al., 2021]	RFC , LR , SVM , NB	RFC =84,80% , LR = 83,82%	Cardiovascular Disease Pre-diction using Machine Learning Algorithms
2022	[Gupta et al., 2022]	KNN, LR , SVM , DT , RF , SVM	LR =92,3%	Prédiction des maladies cardiaques à l'aide de techniques d'apprentissag e automatique supervisé
2023	[Bhatt et al., 2023b]	DT	DT=72.77%	Effective Heart Disease Prediction Using Machine Learning Techniques

TABLE 1.3 – Récapitulation des travaux connexes

Chapitre 2

La conception d'un système intelligent de prédiction des maladies cardiovasculaire

2.1 Introduction

L'apprentissage automatique (machine learning) offre de nombreuses applications dans le traitement des maladies cardiaques. Un des principaux exemples est celui de la prédiction des risques, où les algorithmes d'apprentissage automatique peuvent analyser une grande quantité d'informations sur les patients pour déterminer les paramètres de risque des maladies cardiaques. Cela comprend des facteurs tels que l'âge, le sexe, l'IMC, la tension artérielle, le taux de cholestérol et les antécédents familiaux. Les algorithmes peuvent ensuite utiliser ces différentes informations pour prévoir si un patient est susceptible de souffrir d'une maladie cardiaque à l'avenir. L'apprentissage automatique peut également être utilisé pour améliorer la précision du diagnostic, en analysant les données des patients afin d'identifier des schémas anormaux et de détecter les signes précoces d'une maladie cardiaque. De plus, il peut être utilisé pour optimiser certains schémas diagnostiques, en analysant les données des patients afin d'identifier les traitements les plus efficaces pour certains groupes de patients. Dans l'ensemble, l'apprentissage automatique devrait permettre de faire évoluer la manière dont nous diagnostiquons et traitons les maladies cardiaques, et d'améliorer les résultats pour les patients. Nous décrirons ci-après notre système de prédiction des maladies cardiovasculaires. Ce système repose en effet sur un ensemble de données recueillies auprès d'un important nombre de patients, tout en passant par plusieurs phases. Nous commençons par un prétraitement des données, qui sont ensuite adaptées au processus d'application de l'algorithme d'apprentissage automatique DT . Une fois les données divisées en deux groupes, l'un pour la phase d'apprentissage et l'autre pour la phase de test, le système est entraîné pour obtenir un modèle qui puisse prédire l'état de santé du patient en se basant sur les informations fournies le concernant "Le patient".

2.2 L'architecture détaillée du système

Afin de construire un modèle d'apprentissage automatique pour la classification avec DECISION TREE cela peut être soigneusement classé, nous suggérons le plan dans le schéma *(Figure : 2.1)*.

’

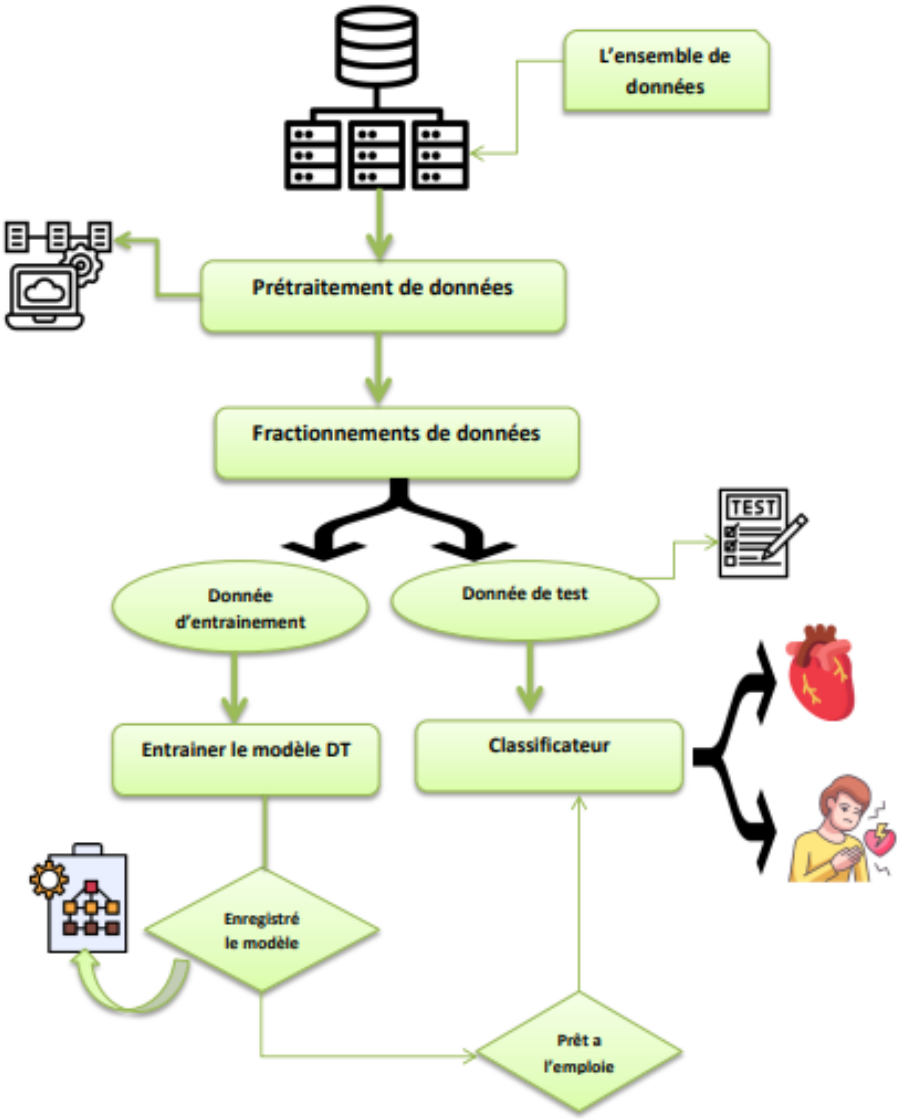


FIGURE 2.1 – Architecture détaillée du système

2.2.1 Lensemble de données

On a utilisé pour ce travail des données médicales provenant d'un ensemble de personnes regroupées dans la base de données Heart Disease Dataset [Smith, 2019], dont nous avons déjà discuté dans le premier chapitre. Il est avantageux d'explorer ces données et de les présenter de manière plus compréhensible en les traduisant dans un format accessible. Cela permet de les interpréter dans un contexte approprié.

2.2.2 Exploration des données

La Figure 2.2 représente les deux classes que contient la base de données, ou (don't have disease) représente la classe non malade et (have disease) représente la classe non malade, et il apparaît également qu'ils sont proches en termes de nombre de patient . De

1025 patients, Il y a 526 patients avec une maladie cardiaque qui est de 51, 32% et il y a 499 patients sans maladie cardiaque qui est de 48, 68% pour cent

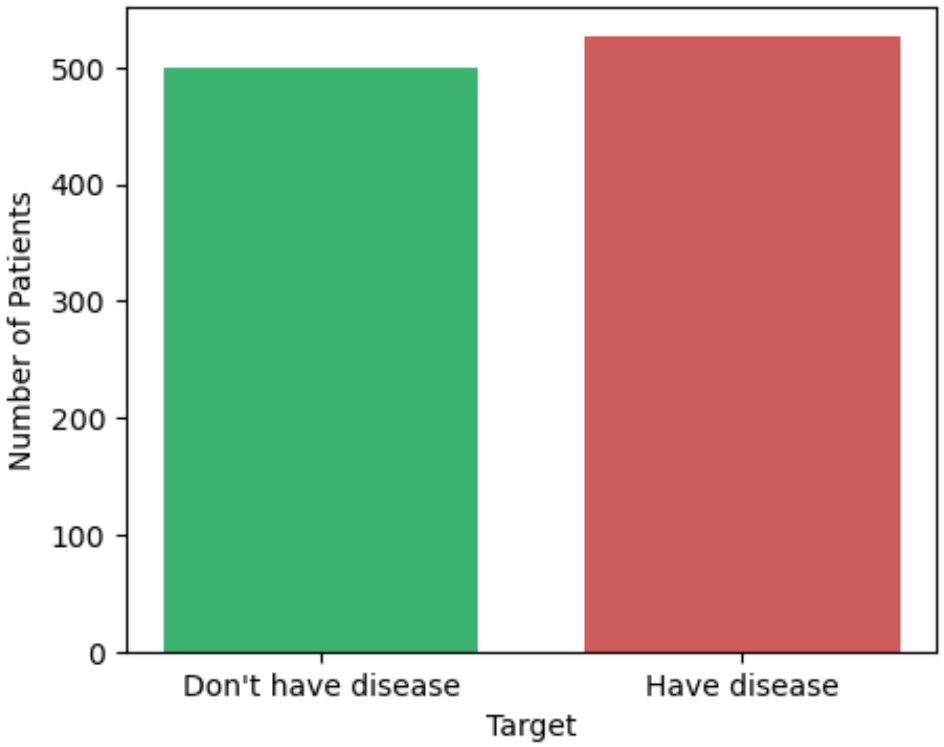


FIGURE 2.2 – Graphe illustrant les deux classes

Un des indicateurs fournissant des informations qui joue un rôle relatif dans la détermination si la maladie est identifiée ou non est l'âge. La Figure 2.3 représente la répartition des patients par âge où les âges varient entre 29 et 77 ans. La différence entre les patients atteints d'une maladie cardiaque et ceux qui n'en sont pas atteints est abondante entre 41 et 45 ans et entre 51 et 54 ans.

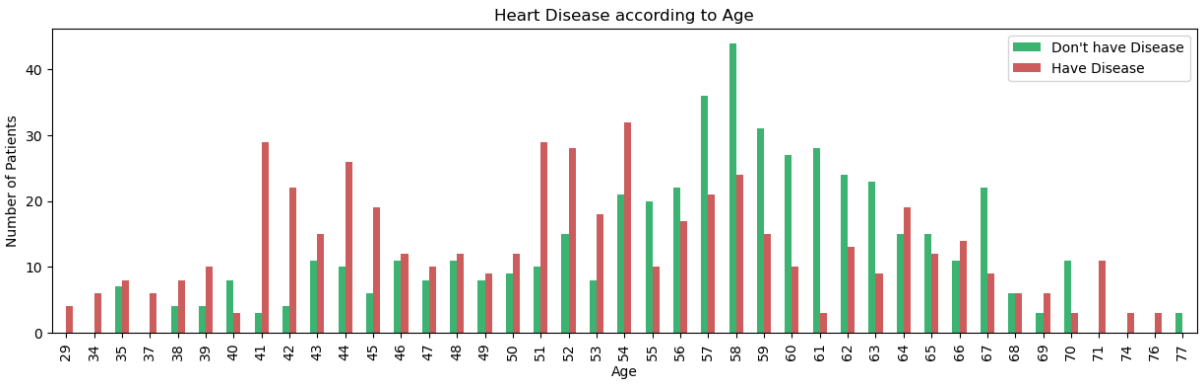


FIGURE 2.3 – Visualisation relation age /maladie cardiaque

Le nombre de femmes atteintes d'une maladie cardiaque est plus élevé que celui des femmes qui n'en ont pas. Par contre, le nombre de patients de sexe masculin atteints d'une maladie cardiaque est inférieur à celui des patients de sexe masculin qui n'ont pas de maladie cardiaque. La figure 2.4 représente la la différence entre les femmes et les hommes.

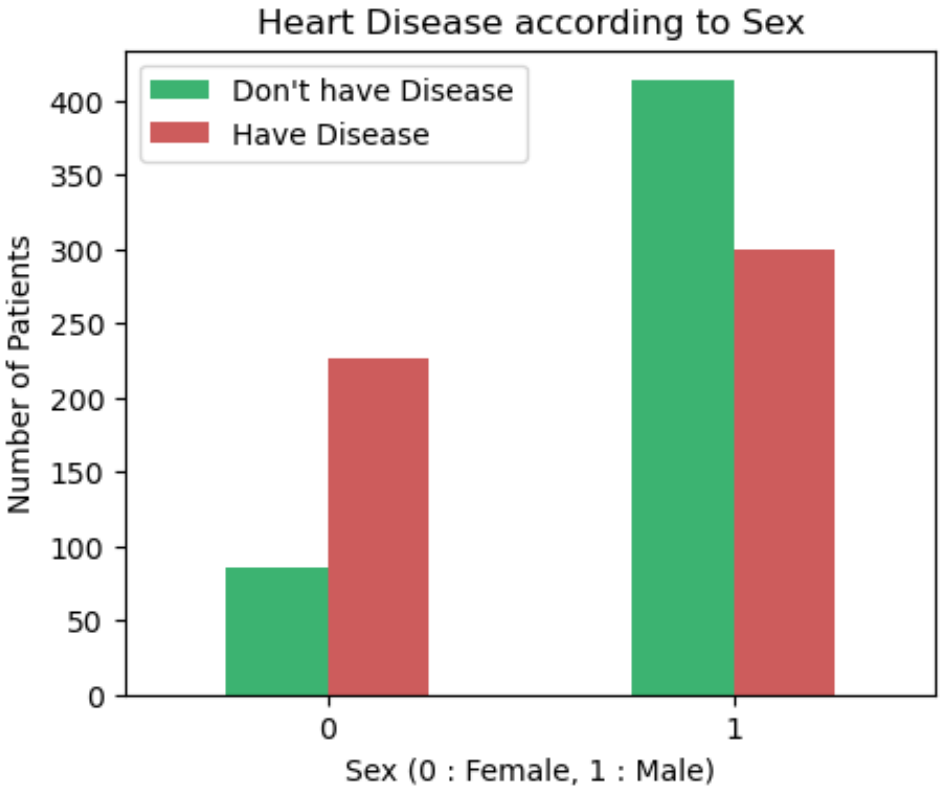


FIGURE 2.4 – – Visualisation relation sex /maladie cardiaque

L'un des processus importants pour améliorer les donées est de déterminer la corrélation entre les variables, ou elle est utilisée pour représenter la mesure statistique de la relation linéaire entre deux variables. Elle peut également etre définie comme une mesure de dépendance entre deux variables différentes.. et pour mieux comprendre ces indicateurs. Nous avons créé une matrice de corrélation La Figure 2.5 qui nous montrerait la relation entre les indicateurs et l'étendue des relations entre eux.
Les valeurs que nous obtenons du coefficient de corr´elation sont limit´ees entre -1 et 1 .
— Si la valeur est -1, nous dirons qu'il s'agit d'une corrélation négative entre deux variables. Cela signifie que lorsqu'une variable augmente, l'autre variable diminue.
— Si la valeur est 0, il n'y a pas de corrélation entre deux variables. Cela signifie que les variables changent de maniére aléatoire les unes par rapport aux autres.
— Si la valeur est 1, nous dirons qu'il s'agit d'une corrélation positive entre deux variables. Cela signifie que lorsqu'une variable augmente, l'autre variable augmente également.
Pour notre cas, nous n'éliminons aucune variable et le jeu de données prétraité final comprenait 14 variables.

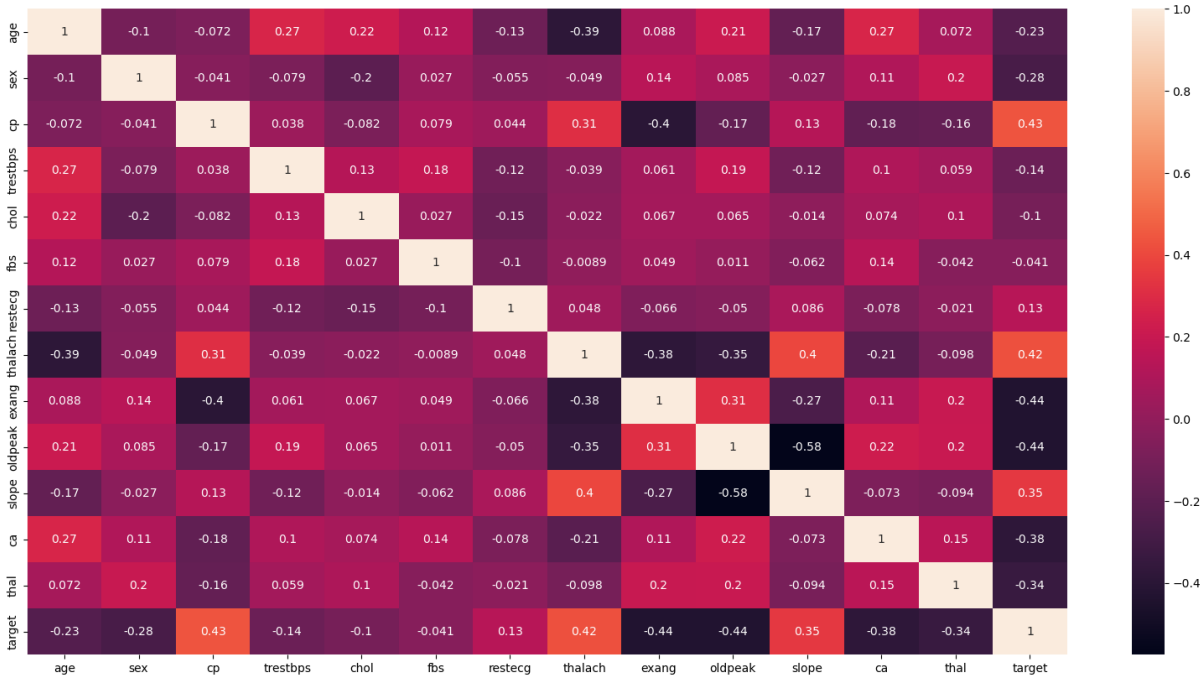


FIGURE 2.5 – Matrice de corrélation entre les indicateurse

La plupart des patients ont un taux de glycémie inférieur à 120 mg/dl, ce qui est normal. Et le nombre de patients avec et sans maladie cardiaque n'est pas différent. Par conséquent, on peut conclure que les taux de sucre dans le sang n'affectent pas le diagnostic de maladie cardiaque. La figure 2.6 représente cette différence :

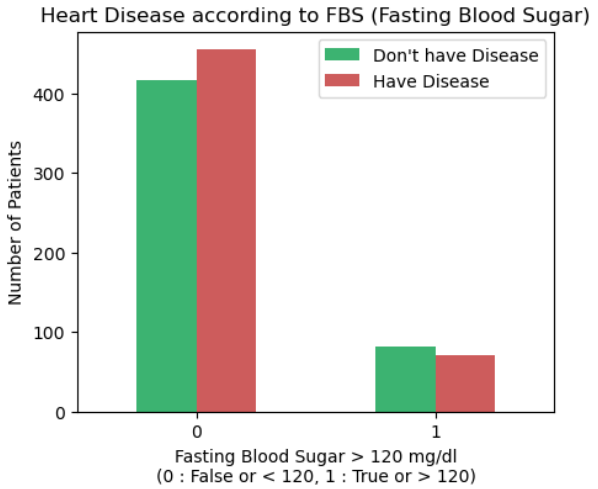


FIGURE 2.6 – – Visualisation relation fbs /maladie cardiaques

2.2.3 Le prétraitement de données

Le prétraitement des données est une étape cruciale lors de la construction d'un modèle d'arbre de décision. Il vise à préparer les données d'entrée de manière à optimiser les performances et la précision du modèle , Tout d'abord, il est essentiel de traiter les valeurs manquantes dans le jeu de données (VOIR la FIGURE 2.7) .

```
In [5]: df.isnull().sum()

Out[5]: age         0
        sex         0
        cp          0
        trestbps    0
        chol        0
        fbs         0
        restecg     0
        thalach     0
        exang       0
        oldpeak     0
        slope       0
        ca          0
        thal        0
        target      0
        dtype: int64
```

FIGURE 2.7 – Les valeurs manquantes

Il est souvent nécessaire de convertir les variables catégorielles en variables numériques , en utilisant des techniques telles que le codage *one-hot*.

De plus, il est recommandé de normaliser ou de standardiser les caractéristiques numériques afin de les mettre à la même échelle. Enfin, il est important de diviser le jeu de données en ensembles d'entraînement et de test pour évaluer les performances du modèle. La proportion de division peut varier en fonction de la taille du jeu de données et des exigences spécifiques du problème. Il y a différentes étapes nécessaires à ce stade pour y parvenir, Parmi eux Voir le figiure 2.8 :

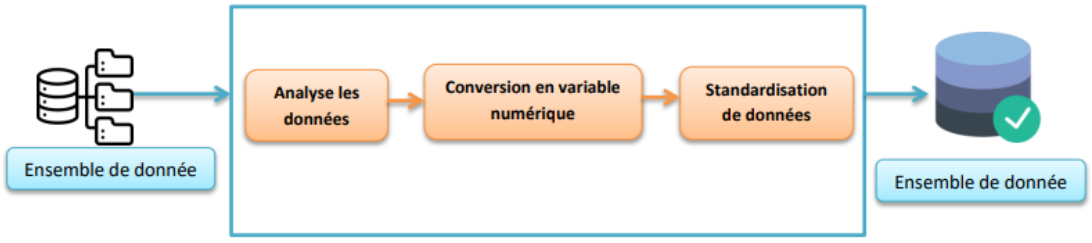


FIGURE 2.8 – Phase de prétraitement

2.2.4 Nettoyage des données

Le nettoyage des données joue un rôle fondamental dans la construction d'un modèle d'arbre de décision précis et fiable. Avant d'entraîner le modèle, il est crucial de procéder à un nettoyage approfondi des données. Ce cas, il est recommandé de détecter et de supprimer ces valeurs aberrantes ou de les remplacer par des valeurs plus appropriées. Lorsque toutes ces opérations de nettoyage sont effectuées, les données sont prêtes à être utilisées pour entraîner un modèle d'arbre de décision .

2.2.5 Standardisation des données

Cette technique permet de s'assurer que les variables ne sont pas biaisées par leur échelle initiale lors de la formation d'un modèle d'apprentissage machine.La standardisation consiste à centrer et réduire les données. Pour centrer les données, on soustrait la moyenne de chaque variable. Pour réduire les données, on divise chaque variable par son écart-type. Ainsi, les données standardisées ont une moyenne de zéro et un écart-type de un. Mais avant cela, nous devons considérer certaines variables catégorielles telles que : sexe, ca, fbs, etc.., ne sont pas comme d'autres variables considérées comme indicatives telles que : chol, trestbps,..,etc. Il faut donc convertir ces variables et puis nous continuons le processus de standardisation, selon la formule suivante :

$$X' = \frac{X - \mu}{\sigma} \tag{2.1}$$

Où μ est la moyenne des valeurs des caractéristiques .
et σ est l'écart type des valeurs descaractéristiques .
X' ce le valeur standarisé .

Et après avoir appliqué la formule à l'ensemble de données, nous obtenons des données standardisées prêtes pour la prochaine étape. Si on l'applique, par exemple, à l'indicateur d'âge, on trouve : Exemple :X= 60

$$X' = \frac{60 - 54.43}{9.07}$$

On a X' = 0.61 $\in$ [0, 1]

2.2.6 Extraction des caractéristiques

L'extraction de caractéristiques consiste à collecter et préparer les données, sélectionner les caractéristiques pertinentes, extraire les caractéristiques à partir des données et évaluer les caractéristiques pour s'assurer qu'elles sont pertinentes et améliorent les performances de prédiction.

Étant donné que l'analyse complexe des données exige une utilisation intensive de la mémoire et du traitement, ou des algorithmes de classification qui exigent un seuil d'ajustement élevé avec les échantillons à tester, elle simplifie le coût des ressources nécessaires pour décrire adéquatement un ensemble de données important. Le résultat de cette étape est de déterminer un sous-ensemble de toutes les caractéristiques initiales, connues sous le nom de caractéristiques extraites.

2.2.7 Le fractionnement de données

Cette technique consiste à diviser un ensemble de données en deux ensemble , un ensemble un ensemble d'apprentissage à 80% et un ensemble de test à 20% (voir la Figure 2.9), L'ensemble d'apprentissage est utilisé pour entraîner le modèle, et l'ensemble de test est utilisé pour évaluer le modèle.

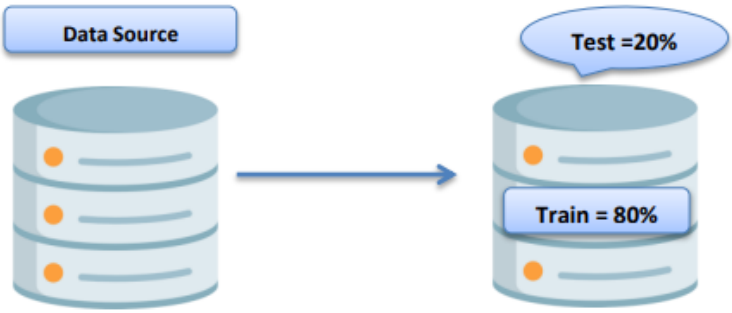


FIGURE 2.9 – Fractionnement de données

2.2.8 Entraîner le modèle

Pour exécuter un modèle d'arbre de décision, séparez l'ensemble de données d'entrée en petits groupes jusqu'à ce que chaque groupe ait exactement une étiquette de sortie. La préparation des données d'entrée est l'étape initiale de l'exécution d'un modèle d'arbre de décision après cela un arbre de décision est créé en utilisant les données du lecteur et l'algorithme de l'arbre de décision. La variable qui maximise la séparation entre les différentes classes de sortie est choisie pour concevoir l'arbre. Pour ce faire, on mesure l'impureté de chaque variable et on choisit celle qui entraîne la plus grande réduction de l'impureté. En comparant les valeurs prévues de la variable cible pour les données d'essai avec les valeurs réelles, la performance du modèle est évaluée.

Les étapes clés d'un modèle d'arbre de décision comprennent la collecte de données d'entrée, l'application de l'algorithme de l'arbre de décision pour créer l'arbre de décision, la prévision des résultats pour de nouveaux exemples et l'évaluation du rendement du modèle.

2.3 Conclusion

Ce chapitre a été consacré à la présentation du système que nous proposons pour prédire les maladies cardiovasculaires, Le système suggéré utilise l'algorithme arbre de décision pour la classification, puis nous avons fini de travailler sur notre projet. Après cela, nous devrons effectuer des tests et évaluer la performance, ce que nous ferons dans le chapitre suivant.

Chapitre 3

Implémentation et résultats expérimentaux

3.1 Introduction

Dans ce chapitre, nous présentons l'environnement de travail, le langage de programmation et tous les outils utilisés dans cette étude Jupter Book, Python et Anaconda,et affichons quelques images de l'interface de l'application et de la base de données que nous avons utilisées en plus d'expliquer la méthode utilisée et les résultats obtenus et comparaison avec un autre modèle.

3.2 Validation du système

Pour valider notre Système on utilise l'ensemble de donnée de test de le même ensemble globale que on a fractionné entre apprentissage avec 80% des données et test avec 20%, la Figure 3.1 affiche la structure de l'ensemble de données globale avec 14 colonnes et 1025 instances.

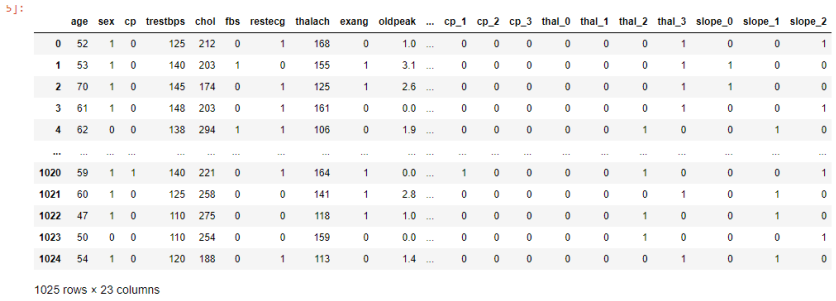
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1020	59	1	1	140	221	0	1	164	1	0.0	2	0	2	1
1021	60	1	0	125	258	0	0	141	1	2.8	1	1	3	0
1022	47	1	0	110	275	0	0	118	1	1.0	1	1	2	0
1023	50	0	0	110	254	0	0	159	0	0.0	2	0	2	1
1024	54	1	0	120	188	0	1	113	0	1.4	1	1	3	0

1025 rows × 14 columns

FIGURE 3.1 – La structure de l'ensemble de données

3.3 Prétraitement des données

Dans un premier temps, nous traitons les variables catégoriques et les transformons des variables indicatrices, ce qui nous permet de généraliser leur traitement à l'aide des mêmes formules et d'obtenir la structure de la figure 3.2 avec 23 colonnes (*on affiche le sommet de l'ensemble de donnée*).



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	...	cp_1	cp_2	cp_3	thal_0	thal_1	thal_2	thal_3	slope_0	slope_1	slope_2
0	52	1	0	125	212	0	1	168	0	1.0	...	0	0	0	0	0	0	1	0	0	1
1	53	1	0	140	203	1	0	155	1	3.1	...	0	0	0	0	0	0	1	1	0	0
2	70	1	0	145	174	0	1	125	1	2.6	...	0	0	0	0	0	0	1	1	0	0
3	61	1	0	148	203	0	1	161	0	0.0	...	0	0	0	0	0	0	1	0	0	1
4	62	0	0	138	294	1	1	106	0	1.9	...	0	0	0	0	0	1	0	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1020	59	1	1	140	221	0	1	164	1	0.0	...	1	0	0	0	0	1	0	0	0	1
1021	60	1	0	125	258	0	0	141	1	2.8	...	0	0	0	0	0	0	1	0	1	0
1022	47	1	0	110	275	0	0	118	1	1.0	...	0	0	0	0	0	1	0	0	1	0
1023	50	0	0	110	254	0	0	159	0	0.0	...	0	0	0	0	0	1	0	0	0	1
1024	54	1	0	120	188	0	1	113	0	1.4	...	0	0	0	0	0	0	1	0	1	0

1025 rows × 23 columns

FIGURE 3.2 – Conversion des variables

Ensuite on applique la formule de standardisation que nous avons présenté dans le chapitre précédent à l'ensemble globale de données, nous obtenons des données normalisées prêtes pour la prochaine étape, comme le montre la figure 3.3 (*le sommet de l'ensemble de donnée*).

Le second résultat représente les caractéristiques après standardisation. Chaque ligne de la matrice correspond à un échantillon ou à une observation, tandis que chaque colonne correspond à une caractéristique spécifique.

En analysant le tableau illustré dans la figure 3.3, on constate que les valeurs ont été ajustées et centrées autour de zéro. Cela signifie que pour chaque caractéristique, la moyenne des valeurs est proche de zéro. De plus, les valeurs sont également mises à l'échelle pour avoir une variance d'environ un.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope
0	52	1	0	125	212	0	1	168	0	1.0	2
1	53	1	0	140	203	1	0	155	1	3.1	0
2	70	1	0	145	174	0	1	125	1	2.6	0
3	61	1	0	148	203	0	1	161	0	0.0	2
4	62	0	0	138	294	1	1	106	0	1.9	1

```
print(X_scaled)
```

```
[[-0.26843658  0.66150409 -0.91575542 ... -0.71228712 -0.06088839
   0.99543334]
 [-0.15815703  0.66150409 -0.91575542 ...  1.40392824  1.72713707
  -2.24367514]
 [ 1.71659547  0.66150409 -0.91575542 ...  1.40392824  1.30141672
  -2.24367514]
 ...
 [-0.81983438  0.66150409 -0.91575542 ...  1.40392824 -0.06088839
  -0.6241209 ]
 [-0.4889957  -1.51170646 -0.91575542 ... -0.71228712 -0.91232909
   0.99543334]
 [-0.04787747  0.66150409 -0.91575542 ... -0.71228712  0.27968789
  -0.6241209 ]]
```

FIGURE 3.3 – Standardisation des caractéristiques

3.4 Le fractionnement de données

Après le prétraitement des données que nous utilisons indiqué dans la sous-section précédente, nous devrions diviser notre ensemble de données. Le fractionnement des données

est un élément crucial de la science des données. La scission fait référence au processus de division données en deux parties ou plus.

Notre ensemble de données est divisé en deux parties, la première est utilisée pour entraîner le modèle avec des données à 80 %, et la seconde est utilisée pour évaluer ou tester les données avec des données à 20 %. Le code ci-après décrit la solution (voir la figure 3.4 :

Splitting training set and test set.

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, random_state=0,test_size=0.2)

print(X_train.shape, X_test.shape, y_train.shape, y_test.shape)
```

```
(820, 11) (205, 11) (820,) (205,)
```

FIGURE 3.4 – Divisez les données

3.5 Test du modèle

Les résultats ont été obtenus en appliquant différents algorithmes de classification à l'ensemble de données utilisé. Nous avons utilisé 5 algorithmes d'apprentissage statistique que nous leur avons comparés, en constatant que le taux de précision (*Accuracy)* dont les résultats sont fourni par la figure 3.5 avec les taux les plus élevés pour KNN, SVM, LR, DT, NB.

```
k-NN test score : 0.8488
Logistic Regression test score : 0.8195
SVM test score : 0.9024
Naive Bayes test score : 0.8195
Decision Tree test score 0.9902
```

FIGURE 3.5 – Les résultats de comparaison

L'algorithme d'arbre de décision s'est avéré être une approche efficace pour la prédiction des maladies cardiovasculaires. Il a démontré une précision élevée (99,02%) comme le montre la figure 3.5, avec une capacité à classer correctement un pourcentage élevé d'échantillons de test.

3.5.1 Matrice de confusion de modèle

La matrice de confusion est un outil essentiel dans l'évaluation de la qualité du modèle et sa capacité de prédiction. Elle offre une vue détaillée des performances de classification du modèle en comparant les prédictions du modèle aux valeurs réelles des données.

Pour les algorithmes que nous avons développés pour l'apprentissage, les matrices de confusion pour les différents modèles sont les suivantes (*voir la figure 3.6*)

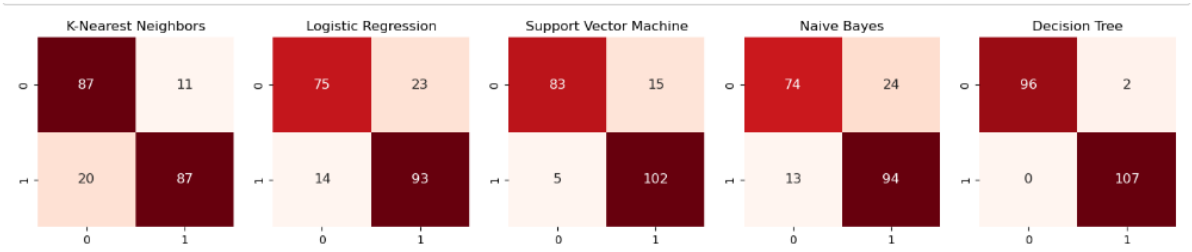


FIGURE 3.6 – Matrices de confusion pour les différents modèles

3.5.2 Les mesures d'évaluation de performance

Les mesures d'évaluation de performance sont des outils essentiels pour évaluer la qualité d'un modèle d'apprentissage automatique. Elles permettent de quantifier et d'analyser les performances du modèle en comparant ses prédictions aux valeurs réelles des données.

* Précision (Accuracy) : La précision mesure la proportion d'échantillons correctement classés par rapport au nombre total d'échantillons. Elle donne une indication globale de l'exactitude du modèle.

* Rappel (Recall) : Le rappel, également connu sous le nom de sensibilité ou taux de vrais positifs, mesure la proportion d'échantillons positifs réellement identifiés par le modèle par rapport au nombre total d'échantillons positifs réels. Il évalue la capacité du modèle à détecter les cas positifs.

* F-mesure (F-measure) : La F-mesure est une mesure qui combine la précision et le rappel en une seule valeur. Elle permet d'obtenir un équilibre entre la précision et le rappel, offrant ainsi une évaluation plus complète des performances du modèle.

Ces mesures d'évaluation de performance permettent d'obtenir des informations précieuses sur les performances du modèle. Elles sont souvent utilisées en conjonction avec d'autres outils d'évaluation tels que la matrice de confusion pour obtenir une vue plus détaillée des performances de classification. Il est important de choisir les mesures d'évaluation appropriées en fonction du problème spécifique et des objectifs du modèle.

Ci-dessous les résultats de chaque algorithme employé (voir les figures [Figure 3.11,Figure 3.7, Figure 3.8, Figure 3.9,Figure 3.10) :

La comparaison de la précision, de la matrice de confusion ainsi que des trois scores permet de constater que l'arbre de décision a la plus forte précision et les scores associés sont les plus élevés. On constate que l'arbre de décision possède la meilleure capacité de prédiction en termes de précision et de score.

3.5.3 Test sur le sous-ensemble de teste

Dans cette section, nous exécutons le modèle de prédiction sur le sous-ensemble de tests avec différentes valeurs de max-depth, où max-depth varie entre 1 et 9, et nous

k-NN

	precision	recall	f1-score	support
Have Disease	0.81	0.89	0.85	98
Don't have Disease	0.89	0.81	0.85	107
accuracy			0.85	205
macro avg	0.85	0.85	0.85	205
weighted avg	0.85	0.85	0.85	205

FIGURE 3.7 – Les mesures d'évaluation de performance du KNN

Logistic Regression

	precision	recall	f1-score	support
Have Disease	0.84	0.77	0.80	98
Don't have Disease	0.80	0.87	0.83	107
accuracy			0.82	205
macro avg	0.82	0.82	0.82	205
weighted avg	0.82	0.82	0.82	205

FIGURE 3.8 – Mesures d'évaluation des performances de LR Logistique Régréssion

SVM

	precision	recall	f1-score	support
Have Disease	0.94	0.85	0.89	98
Don't have Disease	0.87	0.95	0.91	107
accuracy			0.90	205
macro avg	0.91	0.90	0.90	205
weighted avg	0.91	0.90	0.90	205

FIGURE 3.9 – Mesures d'évaluation des performances de SVM

Naive Bayes

	precision	recall	f1-score	support
Have Disease	0.85	0.76	0.80	98
Don't have Disease	0.80	0.88	0.84	107
accuracy			0.82	205
macro avg	0.82	0.82	0.82	205
weighted avg	0.82	0.82	0.82	205

FIGURE 3.10 – Mesures d'évaluation des performances de Naive Bayes

avons noté *l'Accuracy* (exactitude) obtenue pour chaque valeur de max-depth.

Pour évaluer le modèle proposé, on se focalise sur certains critères, à savoir l'exactitude *(Accuracy)*, le rappel, la précision et le F1-score. *Accuracy* est l'une des mesures de per-

```
Decision Tree
--------------------------------------------------------------
                    precision    recall  f1-score   support

      Have Disease       1.00      0.98      0.99        98
Don't have Disease       0.98      1.00      0.99       107

          accuracy                           0.99       205
         macro avg       0.99      0.99      0.99       205
      weighted avg       0.99      0.99      0.99       205

--------------------------------------------------------------
```

FIGURE 3.11 – Mesures d'évaluation des performances de l'arbre de décision

formance les plus importantes pour la classification. Cette évaluation de la performance permet de mieux estimer la qualité du système proposé. Où Le paramètre **"max-depth"** correspond à la profondeur maximale de l'arbre de décision.

Différents accuray avec différents max-depth	
max-depth = 9	Accuray = 99,02 %
max-depth = 8	Accuray = 94,63%
max-depth = 7	Accuray = 93,31%
max-depth = 6	Accuray = 91,71%
max-depth = 5	Accuray = 88,78 %

TABLE 3.1 – Différents accuracy avec différents max-depth du Modéle DT.

La table 3.1 représente les résultats de quatre modèles de l'algorithme **DT**, et dans chaque modèle le paramètre de l'algorithme est modifié pour obtenir une meilleure précision. Où l'on remarque qu'à chaque fois qu'on augmente le paramètre de l'algorithme (*max-depth*) les résultats sont stables et précis.

3.5.4 Test sur un échantillon de teste

Pour tester notre modèle nous avons sélectionné un échantillon de patients (P1,P2,...,P8) de l'ensemble de données afin d'effectuer le test sur eux comme il apparaît dans la Table 3.2, et pour prédire la probabilité des maladies cardiovasculaires.

L'utilisation de la méthode des probabilités pour prédire les maladies cardiovasculaires présente un certain nombre d'avantages, notamment en ce qui concerne l'identification des personnes présentant un risque élevé de maladie cardiovasculaire, avant même l'apparition des symptômes, grâce à des modèles statistiques fondés sur les probabilités. Cela permet de prendre des mesures préventives précoces, telles que des changements de mode de vie ou un traitement médical, afin de réduire le risque de complications graves.

3.6 Discussion des résultats

D'aprés la table 3.1, le modèle de DT atteint une exactitude de 99,02% à la valeur de max-depth= 9 .
Tout d'abord, il convient de souligner que l'exactitude élevée de 99,02% du modèle est

Age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	résultas
52	1	0	125	212	0	1	168	0	1.0	2	00 %
47	1	0	110	275	0	0	118	1	1.0	1	00 %
59	1	1	140	221	0	1	164	1	0.0	2	71.43 %
59	1	2	126	218	1	1	134	0	2.2	1	71.43 %
50	0	0	110	254	0	0	159	0	0.0	2	100 %
71	0	0	112	149	0	1	125	0	1.6	0	100 %

TABLE 3.2 – Résultats de l'échantillon

remarquable et indique une capacité de prédiction extrêmement précise. Cela signifie que le modèle est capable de classer correctement près de la totalité des cas de maladies cardiovasculaires dans notre jeu de données.

En ce qui concerne l'interprétation des résultats, l'utilisation d'un modèle d'arbre de décision offre l'avantage de fournir une visualisation graphique des règles de décision utilisées par le modèle. Cela permet aux praticiens de la santé d'identifier les caractéristiques les plus importantes dans la prédiction des maladies cardiovasculaires et de les utiliser comme base pour de futures recherches ou actions préventives.

En conclusion, notre modèle basé sur un arbre de décision présente une exactitude exceptionnelle de 99,02% dans la prédiction des maladies cardiovasculaires (*voir la table 3.3*).

Modele	Accuracy (%)	Précision (%)	F1-score (%)	Rappel (%)
Decision tree	99,02	100	99	98

TABLE 3.3 – Une évaluation du modèle proposé avec quelques critères de performance.

3.7 Implémentation du système

3.7.1 Préparation de l'environnement de mise en oeuvre

Lors de la préparation de l'environnement implémentation, nous avons installée la distribution Anaconda pour le systèmeème d'exploitation Windows 10 .

3.7.2 Outils et langages de d´eveloppement

Notre choix s'est orienté vers le langage de programmation Python et jupyter , ainsi que des bibliothèques pour la visualisation mathématique et exploration de données (Matplotlib, Numpy, Pandas, sklearn ....).

3.7.3 PYTHON (3.9.16)

Python est le langage parfait pour les scripts et le d´eveloppement rapide dans de nombreux domaines. Il dispose de structures de données de haut niveau efficaces et d'une approche simple mais efficace de la programmation orientée objet. Une syntaxe de serpent

élégante et une écriture dynamique ainsi que sa nature interprétée en font langage d´eal pour les scripts et le d´eveloppement rapide d'applications dans de nombreux domaines sur la plupart des plates-formes. Python Translator et la bibliotheque standard spacieuse sont disponibles gratuitement sous la forme d'une double source pour toutes les principales plates-formes et contiennent de nombreux modules, programmes et outils gratuits Python [Van Rossum and Drake Jr, 1995], Dans notre projet nous avons utilisé la version Python 3.9.16 de Python avec ces bibliothèques.

3.7.4 Pandas (2.0.2)

Pandas est une bibliothèque *Python* sous licence BSD open source fournissant des structures de données et des outils d'analyse de données haute performance et faciles à utiliser pour le langage de programmation Python. Le *Python* avec *Pandas* est utilisé dans un large éventail de domaines, y compris académique et commerciale domaines tels que la finance, l'économie, les statistiques, l'analyse,....,etc [Tutorial, 2006]. C'est pour ces avantages que nous l'avons choisi dans notre projet pour faciliter la manipulation de l'ensemble de donnée que nous avons utilisé sous l'extension (.csv).

3.7.5 Numpy (1.24)

NumPy, qui signifie Numerical Python, est une bibliothèque composée d'objets de tableau multidimensionnel et d'une collection de routines pour traiter ces tableaux. L'utilisation de NumPy permet de faciliter les opérations mathématiques et logiques sur les tableaux, et son rôle est similaire au package précédent, Pandas.
Cependant, pour fournir à la communauté l'accès à de nouvelles technologies exploratoires, NumPy est en transition vers un mécanisme de coordination central qui spécifie une API de programmation de tableaux bien définie et l'envoie, le cas échéant, à des implémentations de tableaux spécialisées [Harris et al., 2020].

3.7.6 Matplotlib (3.7.1)

matplotlib est probablement le paquet Python le plus utilisé pour les graphiques 2D. Il fournit à la fois un moyen très rapide de visualiser les données à partir Python et des chiffres de qualité de publication dans de nombreux formats. Nous allons pour explorer matplotlib en mode interactif couvrant les plus courants affaires [Rougier, 2012].

3.7.7 Pyplot

matplotlib.pyplot est une collection de fonctions de style de commande qui font fonctionner **matplotlib** comme **MATLAB**. Chaque fonction pyplot modifie une figure : par exemple, crée une figure, crée une zone de tracé dans une figure, trace quelques lignes dans une zone de tracé, décore le tracé avec des étiquettes,..., etc [Rougier, 2012].

3.7.8 Seaborn

Seaborn est une bibliothèque pour faire des graphiques statistiques en Python. Il fournit une interface de haut niveau pour matplotlib et s'intègre étroitement avec les structures de données pandas. Fonctions dans les fonds marins bibliothèque exposer une

API déclarative, axée sur les jeux de données qui facilite la traduction des questions sur les données dans des graphiques qui peuvent y répondre [Waskom, 2021].

3.7.9 Scikit-learn

Scikit-learn est un module Python intégrant un large éventail de rythmes algorithme d'apprentissage automatique de pointe pour les problèmes supervisés et non supervisés à moyenne échelle. Cette trousse met l'accent sur l'apprentissage machine pour les non-spécialistes qui utilisent un langage général de haut niveau pour faciliter d'utilisation, les performances, la documentation et la cohérence des API. Il a des dépendances minimales et est distribué sous la licence BSD simplifiée, encourageant son utilisation dans les deux milieux académiques et commerciaux [Pedregosa et al., 2011].

3.7.10 Joblib

Joblib est un ensemble d'outils pour fournir un *pipelining* léger en Python, et en particulier : mise en cache disque transparente des fonctions et réévaluation paresseuse calcul parallèle simple et facile. **Joblib** est optimisé pour être rapide et robuste sur les données volumineuses en particulier et a des optimisations spécifiques pour les tableaux NumPy. Mais nous utilisons à partir de cet ensemble d'outils ce qui est spécial pour le stockage (*Joblib.dump*) et la récupération *(Joblib.load)* des données dans un fichier et en spécifions leur extension (.joblib), afin de stocker notre modèle formé et d'exécuter des tests sur lui [Grisel et al., 2013].

3.7.11 Gardio (3.33.1)

Gradio est une bibliothèque open source développée par l'entreprise Gradio.ai. Elle permet de créer facilement des interfaces utilisateur pour les modèles de *machine learning*. **Gradio** facilite l'interaction avec les modèles en fournissant une interface Web simple et intuitive pour les tester et les visualiser.

Avec **Gradio**, les utilisateurs peuvent charger un modèle pré-entraîné, saisir des données en entrée et obtenir les prédictions correspondantes en temps réel.
Elle offre également des fonctionnalités de personnalisation pour les interfaces utilisateur, telles que des champs de saisie de texte, des boutons, des curseurs, des images, etc [Gradio.ai, 2023].

3.7.12 Anaconda Distribution (2.3.1)

Anaconda est un logiciel libre qui vous fournit une boîte à outils adaptée à la recherche et à la science. L'installation de l'**Anaconda** vous donne accès à différents environnements qui vous permettent de coder en Python ou en R. Ces environnements, aussi appelés environnements de développement intégrés (IDE), sont des plateformes ou des applications qui facilitent grandement le développement de code. Ils servent un rôle similaire aux processeurs de texte comme Microsoft Word, Google Doc et Pages pour écrire du texte, mais en vérité, ils sont tellement plus [Rolon-Mérette et al., 2016].

3.7.13 jupyter notebook (6.4.12)

Jupyter, connu sous le nom IPython Notebook, est un web-based, environnement de développement interac. Initialement développé pour Python, il a depuis élargi pour soutenir plus de 40 autres langages de programmation, y compris **Julia** et **R**.

Jupyter permet d'écrire des ordinateurs portables qui contiennent du texte, du code en direct, des images et des équations. Ces portables peuvent être partagés, et peuvent même être hébergés gratuitement sur GitHub [Bloice and Holzinger, 2016].

3.8 Mode d'utilisation de l'application

Dans notre application, il y a une fenêtre principale (voir les Figures *Fig 3.12 et Fig 3.13*), où l'utilisateur peut remplir un formulaire contenant les informations sur le patient dont nous souhaitons prédire l'état.

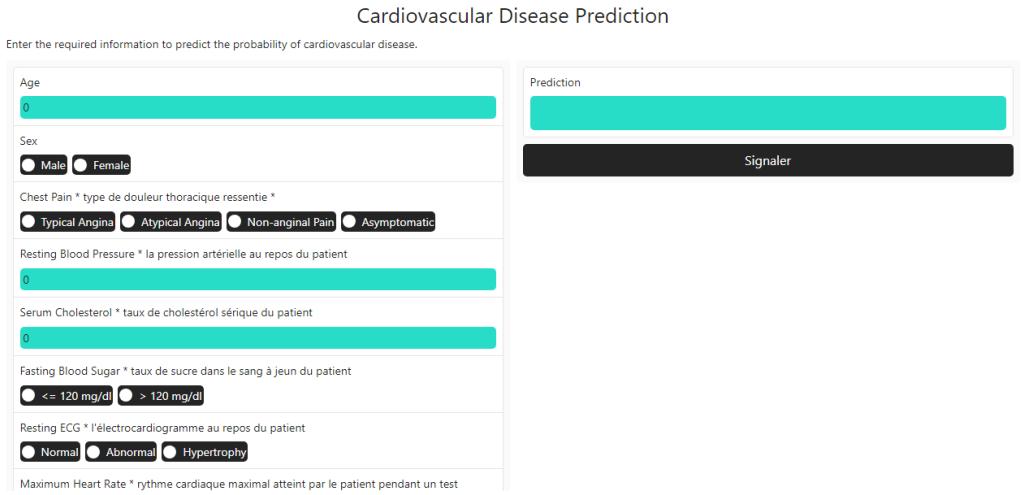


FIGURE 3.12 – la fenêtre principale partie 1

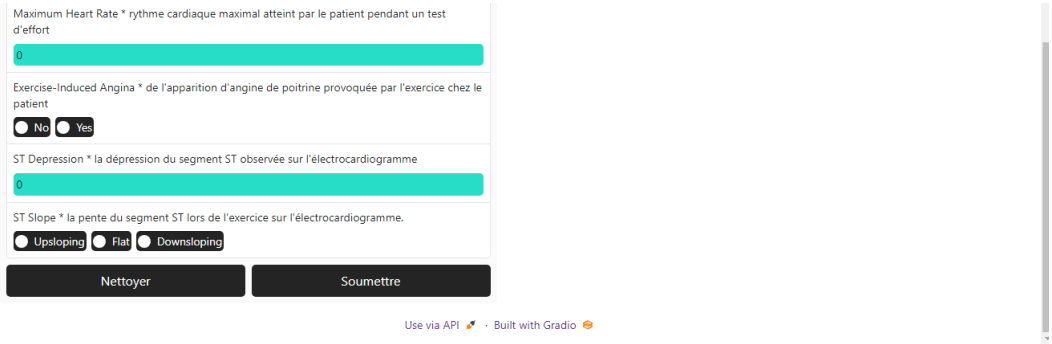


FIGURE 3.13 – La fenêtre principale partie 2

Introduire les informations nécessaires pour prédire la probabilité d'une maladie cardiovasculaire et en cliquant sur le bouton soumettre pour lancer la prédiction, où sur le bouton nettoyer pour vider le formulaire, on obtient un le résultat de la probabilité en matière de maladie cardiovasculaire : voir l'exemple sur le patient qui a eu une probabilité égale à 100% (regarder les figures Fig 3.14 et Fig 3.15) :

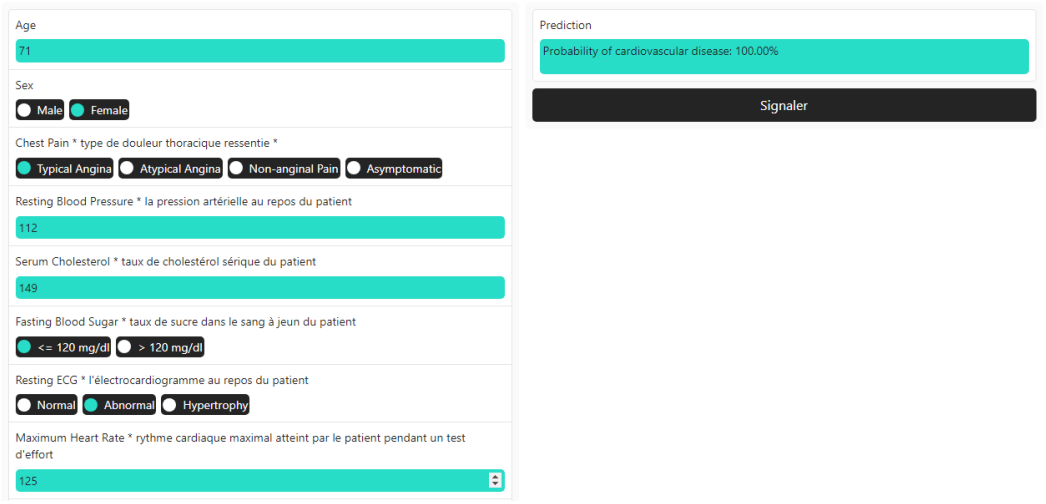


FIGURE 3.14 – Résultas de prédection

et la suite de résultat :

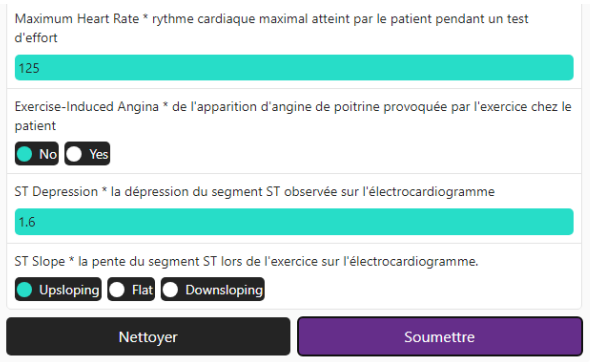


FIGURE 3.15 – Suit Résultas de prédection

Grâce à notre modèle basé sur les probabilités et à notre modèle d'apprentissage DT, il est possible d'identifier les personnes présentant un risque élevé de maladie cardio-vasculaire, avant même l'apparition des symptômes. Cela permet de prendre des mesures préventives précoces, telles que des changements de mode de vie ou un traitement médical, afin de réduire le risque de complications graves.

3.9 Conclusion

Dans ce chapitre, nous avons présenté les différents outils et langage de développement qui permettent l'utilisation d'algorithmes, ainsi que la solution que nous avons proposée pour l'optimisation DT, ou nous avons comparé les résultats de prédiction de certains des algorithmes étudiés avec celle de l'algorithme amélioré. Les résultats expérimentaux obtenus montrent que la démarche proposée apporte une valeur ajoutée au domaine et qu'elle est plus performante que les approches existantes.

Conclusion générale

Cette thèse a réussi à démontrer le potentiel des algorithmes d'apprentissage automatique, en particulier l'arbre de décision, pour améliorer la prédiction des maladies cardiovasculaires. Nous avons ainsi ouvert de nouvelles perspectives de recherche et nous avons souligné l'importance de l'intégration de l'intelligence artificielle dans la pratique médicale.
Dans la suite, nous résumons les contributions que nous avons apportées dans cette thèse et décrivons les principaux axes de ce que nous allons faire dans l'avenir, ainsi que les perspectives ouvertes par l'approche proposée.
Tout d'abord, une étude a été menée pour évaluer et comparer les performances de différents algorithmes d'apprentissage automatique, notamment l'arbre de décision, les K-voisins les plus proches, la machine à vecteur de support et la régression logistique. Cela nous a permis de constater que l'arbre de décision atteignait une précision remarquable de 99,02% dans la prédiction des maladies cardiovasculaires.

Nous suggérons plusieurs directions prometteuses pour nos futurs travaux de recherche. Tout d'abord, il serait intéressant d'explorer plus avant d'autres algorithmes d'apprentissage automatique et d'évaluer leurs performances en matière de prédiction des maladies cardiovasculaires. Cette comparaison approfondie nous permettrait d'identifier l'algorithme le plus performant pour différents contextes cliniques.

En outre, nous recommandons d'enrichir les ensembles de données cliniques utilisés en intégrant des facteurs de risque supplémentaires, tels que les antécédents familiaux, le mode de vie et les données génétiques. Cela permettrait d'améliorer la précision des modèles prédictifs et de mieux comprendre les interactions complexes entre les différents facteurs de risque cardiovasculaire.

Nous encourageons également les collaborations interdisciplinaires entre les chercheurs en informatique de santé et les professionnels de la santé. Ces collaborations permettraient de développer des systèmes intégrés de prédiction des maladies cardiovasculaires, combinant des données cliniques, des données d'imagerie médicale et des informations génétiques, pour une évaluation complète du risque cardiovasculaire.
En poursuivant ce travail à l'avenir, nous pourrons contribuer de manière significative à l'avancement des connaissances et à l'amélioration des soins de santé dans le domaine des maladies cardiovasculaires.

Bibliographie

[Ali et al., 2019] Ali, L., Niamat, A., Khan, J. A., Golilarz, N. A., Xingzhong, X., Noor, A., Nour, R., and Bukhari, S. A. C. (2019). An optimized stacked support vector machines based expert system for the effective prediction of heart failure. *IEEE Access*, 7 :54007–54014.

[Baudet et al., 2012] Baudet, M., Daugareil, C., and Ferrieres, J. (2012). Prévention des maladies cardiovasculaires et règles hygièno-diététiques. In *Annales de Cardiologie et d'Angéiologie*, volume 61, pages 93–98. Elsevier.

[Bhatt et al., 2023a] Bhatt, C. M., Patel, P., Ghetia, T., and Mazzeo, P. L. (2023a). Effective heart disease prediction using machine learning techniques.

[Bhatt et al., 2023b] Bhatt, C. M., Patel, P., Ghetia, T., and Mazzeo, P. L. (2023b). Effective heart disease prediction using machine learning techniques. *Algorithms*, 16(2) :88.

[Bloice and Holzinger, 2016] Bloice, M. D. and Holzinger, A. (2016). A tutorial on machine learning and data science tools with python. *Machine Learning for Health Informatics : State-of-the-Art and Future Challenges*, pages 435–480.

[Crucianu and Philippe, 2010] Crucianu, M. and Philippe, J. (2010). Apprentissage, réseaux de neurones et modèles graphiques.

[Deepika and Sasikala, 2020] Deepika, P. and Sasikala, S. (2020). Enhanced model for prediction and classification of cardiovascular disease using decision tree with particle swarm optimization. In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 1068–1072. IEEE.

[Fayez, 2023] Fayez, M., K. S. (2023). Novel method for diagnosis diseases using advanced high-performance machine learning system.

[Gradio.ai, 2023] Gradio.ai (2023). Gradio : Easy uis for ml models. https://www.gradio.app. Accessed : 07_06_2023.

[Grisel et al., 2013] Grisel, O., Gramfort, A., and Varoquaux, G. (2013). Joblib : Lightweight pipelining in python. *Journal of Machine Learning Research*, 14 :3207–3210.

[Gupta et al., 2022] Gupta, C., Saha, A., Reddy, N. S., and Acharya, U. D. (2022). Cardiac disease prediction using supervised machine learning techniques. In *Journal of Physics : Conference Series*, volume 2161, page 012013. IOP Publishing.

[Harris et al., 2020] Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., et al. (2020). Array programming with numpy. *Nature*, 585(7825) :357–362.

[Jain, 2023] Jain, I. K. H. G. A. T. C. (2023). Fusion of machine learning paradigms : Theory and applications. Intelligent Systems Reference Library. ISBN 978-953-307-035-3 : https ://doi.org/10.1007/978-3-031-22371-6.

[Kumar et al., 2021] Kumar, A., Sharma, G. K., and Prakash, U. (2021). Disease prediction and doctor recommendation system using machine learning approaches. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, 9 :34–44.

[Mahesh, 2020] Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9 :381–386.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn : Machine learning in python. *the Journal of machine Learning research*, 12 :2825–2830.

[Polat et al., 2007] Polat, K., Şahan, S., and Güneş, S. (2007). Automatic detection of heart disease using an artificial immune recognition system (airs) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing. *Expert Systems with Applications*, 32(2) :625–631.

[Prasad Lokulwar, 2022] Prasad Lokulwar, Basant Verma, N. T. K. K. M. B. D. S. (2022). Machine learning methods for engineering application development. Bentham Science Publishers. ISBN : 9815079190,9789815079197.

[Ramalingam et al., 2018] Ramalingam, V., Dandapath, A., and Raja, M. K. (2018). Heart disease prediction using machine learning techniques : a survey. *International Journal of Engineering & Technology*, 7(2.8) :684–687.

[Rivière, 2019] Rivière, D. J.-P. (2019). 10 conseils pour prévenir les maladies cardiovasculaires. https://www.doctissimo.fr/html/sante/mag_2003/special_medec/articlessa_6604_prevention_maladies_cardiovasculaires.htm. Consulté le 20 mai 2023.

[Rolon-Mérette et al., 2016] Rolon-Mérette, D., Ross, M., Rolon-Mérette, T., and Church, K. (2016). Introduction to anaconda and python : Installation and setup. *Quant. Methods Psychol*, 16(5) :S3–S11.

[Rougier, 2012] Rougier, N. P. (2012). *Matplotlib tutorial*. PhD thesis, INRIA.

[Rubini et al., 2021] Rubini, P., Subasini, C., Katharine, A. V., Kumaresan, V., Kumar, S. G., and Nithya, T. (2021). A cardiovascular disease prediction using machine learning algorithms. *Annals of the Romanian Society for Cell Biology*, pages 904–912.

[Saqlain et al., 2016] Saqlain, M., Hussain, W., Saqib, N. A., and Khan, M. A. (2016). Identification of heart failure by using unstructured data of cardiac patients. In *2016 45th International Conference on Parallel Processing Workshops (ICPPW)*, pages 426–431. IEEE.

[Smith, 2019] Smith, J. (2019). Heart disease dataset. https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset. Accessed : Date.

[Son et al., 2010] Son, Y.-J., Kim, H.-G., Kim, E.-H., Choi, S., and Lee, S.-K. (2010). Application of support vector machine for prediction of medication adherence in heart failure patients. *Healthcare informatics research*, 16(4) :253–259.

[Tutorial, 2006] Tutorial, A. B. P. (2006). Tutorial. *Gentry, Apr*, 30.

[Van Rossum and Drake Jr, 1995] Van Rossum, G. and Drake Jr, F. L. (1995). *Python tutorial*, volume 620. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands.

[Waskom, 2021] Waskom, M. L. (2021). Seaborn : statistical data visualization. *Journal of Open Source Software*, 6(60) :3021.

[Weng, 2020] Weng, W.-H. (2020). Machine learning for clinical predictive analytics. *Leveraging data science for global health*, pages 199–217.

[W.h.organisation, 2022] W.h.organisation (2022). Cardiovascular diseases. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1. Consulté le 20 mai 2023.

[Zhang, 2010] Zhang, Y. (2010). Application of machine learning. In-TehOlajnica 19/2, 32000 Vukovar, Croatia. ISBN 978-953-307-035-3.