Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Echahid Cheikh Larbi Tebessi – Tébessa

Faculté des Sciences Exactes et des Sciences de la Nature et de la Vie

Département d'Informatique

Mémoire de fin d'études en vue de

L'obtention du diplôme de MASTER

Domaine : Mathématiques et Informatique

Filière : Informatique

Option : Systèmes d'information

Thème

Real-time smart system for heart disease prediction

Réalisé Par :

Merati isra

Soltani chahinaz

Devant le jury :

• *Hakim Bendjenna*	*Prof.*	*Université Larbi Tébessi- Tébessa*	*Président*
• *Ahmed Zeggari*	*MCA*	*Université Laarbi Tébessi- Tébessa*	*Examinateur*
• *Mohamed Yassine Haouam*	*MCA*	*Université Larbi Tébessi- Tébessa*	*Encadrant*
• *Boudjemaa Khelifa*	*MAA*	*Université Larbi Tébessi- Tébessa*	*Co-Encadrant*

Date de soutenance : 10/06/2024

Remerciement

Nous voulons exprimer en ces quelques lignes de remerciements nos gratitudes envers tout d'abord notre encadrant,

Dr Mohamed Yassine Haouam et **Dr. Boudjemaa Khelifa** pour leurs conseils et leur soutien.

Après à tous ceux en qui présence, leur soutien, leur disponibilité et leurs conseils nous avons trouvé courage afin d'accomplir ce projet.

Nous voulons aussi remercier les membres du jury

Mr. **Ahmed Zeggari** et Mr. **Hakim Bendjenna**, d'avoir accepté d'étudier et évaluer notre travail.

Un respect et un remerciement particuliers à **Mr. Fathi Hamidane**, chef de département, ainsi qu'à tous les enseignants du département d'informatique pour leur précieuse contribution à ce travail.

Merci à tous.

Dédicace

Je dédie avant tout ce succès à mes parents,

El Fajri et **Bounif Naïma**, qui ont beaucoup, pas peu, contribué à m'encourager moralement avant financièrement. Je leur donne toute mon appréciation et mon respect et tout cet effort et tout cet effort. L'avantage d'atteindre ce stade est leur confiance, leurs encouragements et leur confiance en moi.

Que mon Seigneur me les protège et, si Dieu le veut, mon Seigneur me permettra de les soutenir dans leur vieillesse. J’offre ma dévotion à mes sœurs **Douha** et **loujain** , mes frères **Abd rahim , Moataz bllah , Mouaid bllah** et tous les membres de ma famille qui ont contribué à mon encouragement.

Qu'il s'agisse de mes oncles, tantes, oncles ou arrière-grands-mères, j'apprécie leur aide et tout ce qu'ils m'ont apporté.

MERATI isra

Dédicace

A ma très chère mère **Dj SOLTANI**

Quoi que je fasse ou que je dise, je ne saurai point te remercier comme il se doit.

Ton affection me couvre, ta bienveillance me guide et ta présence à mes côtés a toujours été ma source de force pour affronter les différents obstacles.

A mon très cher père **El hocine SOLTANI**

Tu as toujours été à mes côtés pour me soutenir et m'encourager. Que ce travail traduit ma gratitude et mon affection.

A mes très chers frères **Hakim et Hazem** et mes belles sœurs **Hanan**, **Sihem**, **Khaoula**, **Racha**

Puisse Dieu vous donne santé, bonheur, courage et surtout réussite

SOLTANI chahinaz

Table de Matière

Liste des tableaux

Liste des figures

Résumé

Résumé

Les maladies cardiaques sont l'une des principales causes de décès à travers le monde. Malgré cela, prédire ces maladies reste difficile pour les médecins en raison de leur complexité et des coûts élevés associés. C'est pourquoi, ces dernières années, de nombreux chercheurs se tournent vers l'utilisation des technologies modernes telles que l'intelligence artificielle et l'apprentissage automatique pour anticiper ces maladies avant qu'elles ne surviennent.

Le but de cette étude est de proposer un système temps réel pour la prédiction des maladies cardiaques en se basant sur des algorithmes d'apprentissage supervisé et la base de données UCI Cleveland. Nous avons utilisé des algorithmes tels que Machines à Vecteurs de Support (SVM), Forêt d'Arbres Décisionnels (RF), Arbre de Décision (DT), k-plus Proches Voisins (k-NN). Ensuite, nous avons identifié l'algorithme qui a montré le taux d'exactitude le plus élevé, puis nous avons sélectionné les caractéristiques les plus importantes en utilisant l'algorithme génétique pour améliorer les performances et réduire les effets néfastes des caractéristiques non pertinentes et redondantes.

Mots clés : les maladies cardiaques, système temps réel, prediction, apprentissage supervisé, UCI cleveland, Machines à Vecteurs de Support (SVM), Forêt d'Arbres Décisionnels (RF), Arbre de Décision (DT), k-plus Proches Voisins (k-NN), selection des caractéristiques.

Abstract:

Heart diseases are one of the leading causes of death worldwide. Despite this, predicting these diseases remains challenging for doctors due to their complexity and the high associated costs. This is why, in recent years, many researchers have turned to modern technologies such as artificial intelligence and machine learning to anticipate these diseases before they occur.

The objectif of this study is to propose a real-time smart system for predicting heart diseases using supervised learning algorithms and the data available in the UCI Cleveland database. Algorithms such as Support Vector Machines (SVM), Random Forest (RF), Decision Tree (DT), and k-Nearest Neighbors (k-NN) are employed. Subsequently, we identified the algorithm that showed the highest accuracy rate and selected the most important features using the genetic algorithm to improve performance and reduce the detrimental effects of irrelevant and redundant features.

Keywords: heart diseases, real-time smart system, prediction, supervised learning, UCI Cleveland, Support Vector Machines (SVM), Random Forest (RF), Decision Tree (DT), k-Nearest Neighbors (k-NN), feature selection.

ملخص

أمراض القلب هي واحدة من الأسباب الرئيسية للوفاة في جميع أنحاء العالم. وعلى الرغم من ذلك، يبقى التنبؤ بهذه الأمراض تحديًا للأطباء بسبب تعقيدها والتكاليف المرتفعة المرتبطة بها. لهذا السبب، في السنوات الأخيرة، اتجه العديد من الباحثين إلى استخدام التقنيات الحديثة مثل الذكاء الاصطناعي والتعلم الآلي للتنبؤ بهذه الأمراض قبل حدوثها.

تهدف هذه الدراسة إلى اقتراح نظام في الوقت الحقيقي للتنبؤ بأمراض القلب باستخدام خوارزميات التعلم الالي و البيانات المتاحة في قاعدة بيانات .UCI Cleveland تم استخدام خوارزميات مثل آلات الاشعة الحاملة (SVM) ، الغابة العشوائية(RF) ، شجرة القرار (DT)، وأقرب الجيران .(k-NN) بعد ذلك، قمنا بتحديد الخوارزمية التي أظهرت أعلى معدل دقة واختيار الخصائص الأكثر أهمية باستخدام الخوارزمية الجينية لتحسين الأداء وتقليل التأثيرات الضارة للخصائص غير ذات الصلة والمتكررة.

الكلمات المفتاحية : أمراض القلب، نظام في الوقت الحقيقي، التنبؤ، التعلم تحت الإشراف، UCIكلفلاند، الة دعم المتجهات (SVM) ، الغابة العشوائية (RF) ، شجرة القرار (DT) ، اقرب الجيران (K-NN) ،اختيار السمات.

Introduction générale

Le cœur, cet organe vital dans le corps humain, ne se limite pas à simplement pomper le sang à travers tout le corps, mais il représente également un centre vital d'activité biologique et émotionnelle pour l'être humain. Le cœur régule ses battements de manière précise et régulière pour assurer la circulation du sang et nourrir les tissus et organes en oxygène et en éléments nutritifs essentiels à la survie du corps.

Cependant, malgré son importance capitale, le cœur est sujet à de nombreux problèmes et maladies qui peuvent affecter négativement sa fonction et la santé de l'être humain en général. Parmi ces problèmes, on trouve les diverses maladies cardiaques telles que les maladies coronariennes, l'angine de poitrine, les crises cardiaques, l'insuffisance cardiaque, les inflammations cardiaques, les fuites valvulaires, et bien d'autres.

Les maladies cardiaques figurent parmi les principales causes de décès dans le monde, ce qui rend leur prévention et leur diagnostic précoce essentiels pour préserver la santé de l'homme et réduire les risques de les contracter. Ce qui rend la situation plus complexe, c'est que de nombreuses maladies cardiaques peuvent ne pas présenter clairement leurs symptômes aux premiers stades, ce qui rend leur prédiction un défi majeur pour les médecins et les professionnels de la santé.

Ainsi, l'utilisation de technologies telles que l'intelligence artificielle et l'apprentissage automatique peut jouer un rôle crucial dans l'amélioration de la capacité des médecins à diagnostiquer et à prédire les maladies cardiaques avec une précision accrue, améliorant ainsi les chances de traitement et de prévention précoce de ces maladies importantes.

Dans cette étude, nous avons créé un système en temps réel pour prédire les maladies cardiaques en utilisant une approche à deux étapes. La première étape consistait à appliquer des algorithmes de machine learning supervisé tels que Machines à Vecteurs de Support (Support Vector Machines SVM), Forêt d'Arbres Décisionnels (Random Forest RF), Arbre de Décision (Decision Tree DT), k-plus Proches Voisins (k-Nearest Neighbors k-NN), et nous avons choisi le meilleur model qui donné les meilleurs résultats en termes de précision et d'exactitude. La deuxième étape impliquait la sélection des caractéristiques les plus influentes de la base de données en utilisant un algorithme génétique.

Nous avons utilisé la base de données UCI Cleveland Heart Disease Dataset, car c'est la plus utilisée dans ce domaine de recherche.Le reste du manuscrit est structuré comme suit :

- Chapitre 1 s'intitule "Généralités sur les maladies cardiaques" Dans le premier chapitre, nous présentons des informations complètes sur les maladies cardiaques, en commençant par leur définition et en détaillant la structure et le fonctionnement du cœur. Les différentes classifications des maladies cardiaques sont abordées, ainsi que les symptômes associés à chaque type, et les méthodes de diagnostic disponibles pour identifier précisément les affections cardiaques. De plus, le chapitre examine les types de bases de données utilisées dans l'étude des maladies cardiaques et leur importance dans la documentation et l'analyse des données. Les différentes méthodes de traitement disponibles pour différents types de bases de données sont également examinées. Enfin, les principales recherches et études sur le même sujet sont présentées.
- Chapitre 2 "Prédiction des maladies cardiaques par les techniques d'apprentissage automatique "Dans ce chapitre, nous mettons l'accent sur la base de données que nous avons utilisée, la base de données "UCI Cleveland". Nous avons analysé ces données et les avons traitées en utilisant les méthodes appropriées pour les préparer à la phase d'application. Dans la deuxième phase, nous avons appliqué des algorithmes d'apprentissage automatique pour prédire les maladies cardiaques en utilisant les données que nous avions préparées précédemment.
- Chapitre 3 "Sélection des caractéristiques par application Algorithme Génétique "Dans ce dernier chapitre, après avoir choisi l'algorithme avec lequel nous travaillerons, nous avons exploré divers algorithmes pour sélectionner les caractéristiques pertinentes. Nous avons opté pour l'algorithme génétique pour extraire les caractéristiques importantes nécessaires à la prédiction des maladies cardiaques.

Chapitre 1 :

Généralité sur les maladies cardiaques et avancées en Apprentissage Automatique

1. Introduction

Les maladies cardiovasculaires représentent l'une des principales causes de décès dans le monde. Elles englobent un large éventail de pathologies affectant le cœur et les vaisseaux sanguins, telles que l'athérosclérose, l'hypertension artérielle, les troubles du rythme cardiaque, les crises cardiaques et les accidents vasculaires cérébraux. Ces affections constituent un défi majeur pour la santé publique en raison de leur prévalence élevée et de leurs conséquences graves.

Dans ce chapitre, nous abordons divers aspects du cœur et des maladies cardiaques, ainsi que leur prédiction à l'aide des techniques d'intelligence artificielle et d'apprentissage automatique. Nous commencerons une définition du cœur et des maladies cardiaques, en incluant leurs types, leurs symptômes et les facteurs de risque associés. Ensuite, nous explorerons en détail les concepts clés de l'apprentissage automatique, de l'intelligence artificielle et de l'apprentissage profond, ainsi que leur application dans le domaine de la cardiologie. Nous examinerons également les différentes sources de données utilisées dans l'étude des maladies cardiaques et les méthodes de prétraitement avant l'application des techniques d'apprentissage automatique. Enfin, nous passerons en revue les recherches antérieures qui ont contribué à ce domaine, nous permettant de saisir les développements actuels et les tendances futures dans l'étude des maladies cardiaques grâce aux techniques d'apprentissage automatique et d'intelligence artificielle.

2. Concepts généraux et définitions

2.1 Le cœur

Le **cœur** est un muscle creux, situé au niveau du thorax entre les poumons et reposant sur le diaphragme. Ce muscle est une pompe ayant pour fonction de propulser le sang vers tous les organes de l'organisme.

Il pompe 5 litres de sang par minute et battra environ 3 milliard de fois au cours d'une vie entière [1].

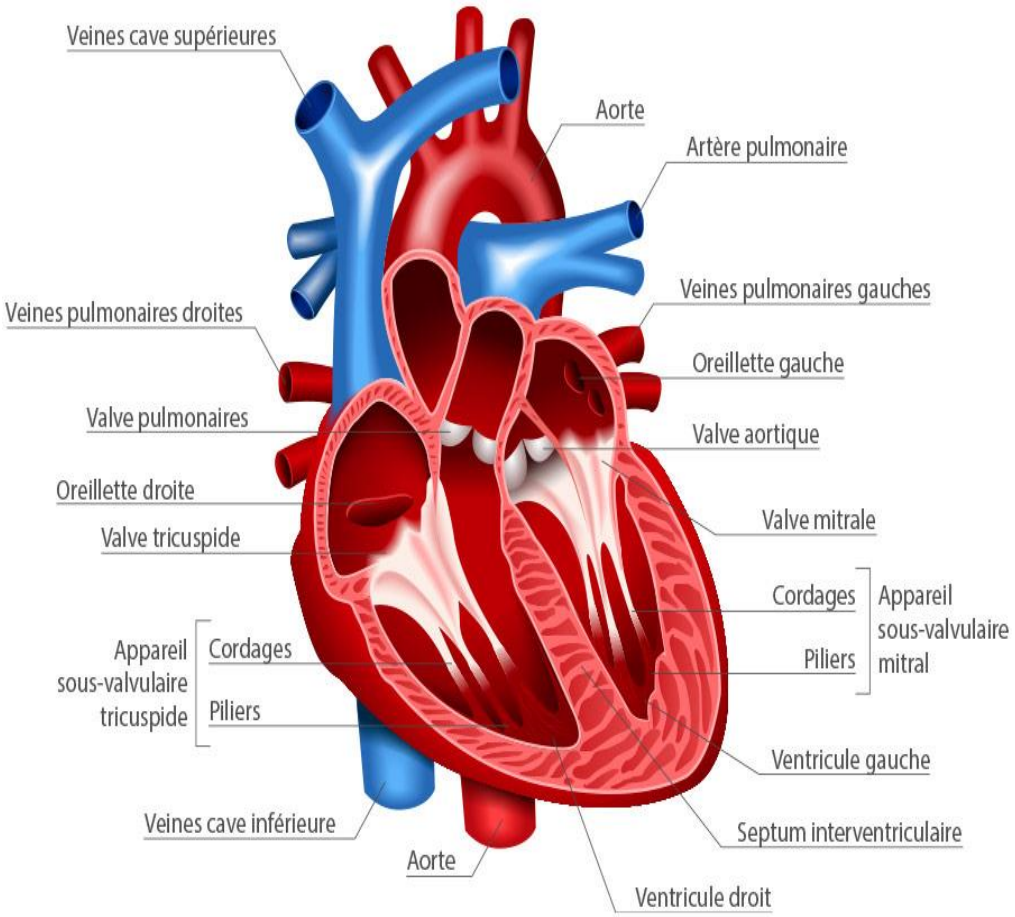

Figure 1 Structure anatomique du cœur. [1]

2.2 Les maladies cardiaques

Les maladies cardiovasculaires désignent l'ensemble des maladies du cœur et des artères. Elles sont causées pour une large part par des dépôts de cholestérol sur les parois des artères. Ces dépôts finissent par gêner, voire empêcher la circulation du sang qui alimente le cœur, le cerveau ou les jambes, entraînant des symptômes tels que les angines de poitrine, les infarctus, les accidents vasculaires cérébraux (AVC) et les artérites [2].

2.3 Types des maladies cardiaques

Les maladies cardiovasculaires constituent un ensemble de troubles affectant le cœur et les vaisseaux sanguins, qui comprennent [3] :

- les cardiopathies coronariennes : Elles affectent les vaisseaux sanguins qui nourrissent le muscle cardiaque.
- les maladies cérébro-vasculaires : Elles touchent les vaisseaux sanguins qui approvisionnent le cerveau.
- les artériopathies périphériques : Elles concernent les vaisseaux sanguins qui alimentent les bras et les jambes.
- les cardiopathies rhumatismales : Elles influent sur le muscle cardiaque et les valves, résultant souvent d'un rhumatisme articulaire aigu causé par une bactérie streptocoque.
- les malformations cardiaques congénitales : Ce sont des anomalies de la structure cardiaque présentes dès la naissance.
- Les thromboses veineuses profondes et les embolies pulmonaires : Il s'agit d'une obstruction des veines des membres inférieurs par un caillot sanguin, pouvant se détacher et migrer vers le cœur ou les poumons.

2.4 Les symptômes des maladies cardiaques

Les symptômes des maladies cardiaques dépendent du type de maladie cardiaque [4] :

2.4.1 Symptômes de maladie cardiaque dans les vaisseaux sanguins :

Les symptômes de la maladie des vaisseaux sanguins peuvent inclure :

- Douleur thoracique, oppression thoracique, pression thoracique et inconfort thoracique (angine).
- Essoufflement.
- Douleur dans le cou, la mâchoire, la gorge, le haut du ventre ou le dos.
- Douleur, engourdissement, faiblesse ou froideur dans les jambes ou les bras si les vaisseaux sanguins de ces zones du corps sont rétrécis.

2.4.2 Symptômes de maladie cardiaque causés par des battements cardiaques irréguliers (arythmies cardiaques) :

Le cœur peut battre trop vite, trop lentement ou de manière irrégulière. Les symptômes de l'arythmie cardiaque peuvent inclure :

- Douleur ou inconfort thoracique.
- Vertiges.
- Évanouissement (syncope) ou quasi-évanouissement.
- Flottement dans la poitrine.
- Étourdissements.
- Rythme cardiaque accéléré (tachycardie).
- Essoufflement.
- Rythme cardiaque lent (bradycardie).

2.4.3 Symptômes de maladie cardiaque causés par des malformations cardiaques congénitales

Les malformations cardiaques congénitales graves sont généralement constatées peu de temps après la naissance. Les symptômes de malformation cardiaque congénitale chez les enfants peuvent inclure :

- Peau ou lèvres gris pâle ou bleues (cyanose).
- Gonflement des jambes, du ventre ou du contour des yeux.
- Chez un nourrisson, essoufflement lors des tétées, entraînant une faible prise de poids.

2.4.4 Symptômes de maladie cardiaque causés par une maladie du muscle cardiaque (cardiomyopathie)

Les premiers stades de la cardiomyopathie peuvent ne pas provoquer de symptômes visibles. À mesure que la maladie s'aggrave, les symptômes peuvent inclure :

- Vertiges, étourdissements et évanouissements.
- Fatigue.
- Se sentir essoufflé pendant une activité ou au repos.
- Se sentir essoufflé la nuit en essayant de dormir ou en se réveillant essoufflé.
- Battements de cœur irréguliers qui semblent rapides, battants ou palpitants.
- Jambes, chevilles ou pieds enflés.

2.4.5 Symptômes de maladie cardiaque causés par des problèmes de valvules cardiaques (cardiopathie valvulaire)

La cardiopathie valvulaire est également appelée maladie des valvules cardiaques. Selon la valvule qui ne fonctionne pas correctement, les symptômes de la valvulopathie cardiaque comprennent généralement :

- Douleur thoracique.
- Évanouissement (syncope).
- Fatigue.
- Rythme cardiaque irrégulier.
- Essoufflement.
- Pieds ou chevilles enflés.

2.5 Les facteurs de risques des maladies cardiaques

Ces facteurs sont multiples. Ils sont classés en facteurs non modifiables et facteurs modifiables [5] :

2.5.1 Facteurs de risque non modifiables

- **Âge et sexe :** Le nombre absolu de décès cardiovasculaires est plus important chez les femmes (54 %) que chez les hommes, mais avant 65 ans la mortalité cardiovasculaire des hommes est 3 à 4 fois supérieure à celle des femmes. En pratique, les accidents cardiovasculaires surviennent en moyenne dix ans plus tôt chez l'homme que chez la femme [16].
- **Hérédité :** Son évaluation repose sur la notion d'événements précoces chez les parents ou dans la fratrie. Ceux-ci peuvent être liés à la transmission génétique de facteurs de risque modifiables (hypercholestérolémie familiale, HTA, diabète…). Mais, c'est souvent la seule présence de facteurs environnementaux familiaux défavorables (tabagisme, alimentation déséquilibrée, sédentarité…) qui explique les accidents sur plusieurs générations.

2.5.2 Facteurs de risque modifiables

- **Tabagisme** : le tabagisme est le deuxième facteur de risque d'Infarctus du Myocarde (IDM), juste derrière les dyslipidémies. Celle-ci confirme que le risque d'infarctus du

myocarde est proportionnel à la consommation, mais il n'y a pas de seuil de consommation au-dessous duquel le tabagisme est dénué de risque. La part attribuable au tabagisme dans la survenue d'un IDM est d'autant plus importante que les sujets sont jeunes. C'est le facteur essentiel et souvent isolé des accidents coronaires aigus des sujets jeunes ; le risque d'IDM concerne également le tabagisme passif. Par ailleurs, à côté des complications coronaires, le tabagisme joue un rôle majeur dans la survenue et l'évolution de l'artériopathie oblitérant des membres inférieurs : 90 % des patients ayant cette localisation d'athérosclérose sont fumeurs. Le risque de développer un anévrisme de l'aorte abdominale est significativement augmenté chez les fumeurs. Enfin, les études épidémiologiques montrent une corrélation entre la consommation de tabac et le risque d'accident vasculaire cérébral aussi bien chez l'homme que chez la femme.

- **Hypercholestérolémie :**C'est le facteur de risque le plus important pour la maladie coronarienne. La cholestérolémie totale est corrélée positivement et de façon exponentielle avec le risque coronaire. Au niveau individuel, le facteur déterminant du risque est un niveau élevé de LDL-cholestérol (cholestérol transporté par les lipoprotéines de basse densité). De façon indépendante, un faible niveau de HDL-cholestérol (cholestérol transporté par les lipoprotéines de haute densité) est un autre facteur de risque de maladie coronarienne, tandis qu'un niveau élevé de HDL-cholestérol est au contraire protecteur. D'où les termes communément utilisés de «mauvais» (LDL-C) et de «bon» cholestérol (HDL-C). Pour la majorité des hypercholestérolémies, les facteurs en cause sont alimentaires et liés à des apports trop importants en acides gras saturés. Mais certaines hypercholestérolémies sont dépendantes de facteurs génétiques.
- **Hypertension artérielle :** L'hypertension artérielle (HTA) est définie par des chiffres tensionnels supérieurs à 140/90 mmHg. Plus la pression artérielle augmente, plus le risque cardiovasculaire est important. Même à des niveaux de pression artérielle inférieurs, le risque de maladies cardiovasculaires reste proportionnel au niveau tensionnel. Aussi L'HTA est le plus souvent silencieuse, avec peu ou pas de symptômes. Elle retentit principalement sur trois organes : le cœur (insuffisance coronaire et insuffisance cardiaque), le cerveau (accident vasculaire cérébral [AVC]) et les reins (insuffisance rénale). La pression artérielle augmente avec l'âge : il y a plus de 50 % d'hypertendus après 65 ans.
- **Diabète :** Le diabète de type 1, insulinodépendant (10 à 15 % des diabétiques), qui débute le plus souvent avant l'âge de 20 ans, est associé à une augmentation significative

du risque cardiovasculaire. Mais c'est surtout le diabète de type 2, non insulinodépendant (85 à 90 % des diabétiques), qui du fait de sa prévalence importante et croissante est dominant dans le risque cardiovasculaire. Le diabète est à l'origine de complications macrovasculaires (maladie coronaire, AVC, artériopathie oblitérant des membres inférieurs [AOMI]).

En plus des facteurs mentionnés précédemment, d'autres éléments contribuent aux maladies cardiaques, tels que l'obésité abdominale et des facteurs psychosociaux.

2.6 Méthode de diagnostic des maladies cardiaques

Il existe de plusieurs méthodes pour diagnostiquer les maladies cardiaques, parmi les plus importantes, nous trouvons [6] :

2.6.1 Électrocardiogramme (ECG)

Un électrocardiogramme (ECG) est un test qui étudie le fonctionnement du cœur en mesurant son activité électrique. À chaque battement cardiaque, une impulsion électrique (ou « onde ») traverse le cœur. Cette onde fait contracter le muscle cardiaque afin qu'il expulse le sang du cœur.

2.6.2 Échocardiogramme

Un échocardiogramme utilise des ondes sonores (ultrasons) pour tracer une image de votre cœur. Les ondes enregistrées montrent la forme, la texture et le mouvement des valvules, ainsi que le volume et le fonctionnement des cavités cardiaques.

2.6.3 Cathétérisme cardiaque

Le cathétérisme cardiaque est une méthode d'exploration utilisée pour effectuer divers tests et interventions. En général, on y a recours conjointement avec d'autres tests tels que l'angiographie et l'étude électrophysiologique.

Le cathétérisme cardiaque permet habituellement de mesurer la pression dans les cavités cardiaques et le fonctionnement du mécanisme responsable de l'apport sanguin, de même que pour dépister des anomalies cardiaques chez les nouveau-nés. Cet examen peut également être utile pour déterminer si une intervention à cœur ouvert s'impose ou

non. Par ailleurs, on peut faire appel à cette méthode dans un but thérapeutique, notamment pour le traitement ou la réparation de malformations cardiaques, pour dilater une valvule sténose ou désobstruer une artère ou un cœur greffé.

2.6.4 Échographie Doppler

Semblable à l'échocardiogramme, l'échographie Doppler est un examen au cours duquel des ondes sonores, émises à très haute fréquence, rebondissent sur votre cœur et vos vaisseaux sanguins. Les ondes, qui sont alors réfléchies tel un écho, sont recueillies et transformées en images qui reflètent la circulation sanguine à travers le cœur et les vaisseaux sanguins. L'échographie Doppler permet donc aux médecins d'examiner, de façon distincte, les conditions d'écoulement et d'irrigation du cœur et des vaisseaux sanguins. Elle leur permet également d'observer et de mesurer les éventuelles obstructions des artères, de même que le degré de rétrécissement ou d'écoulement des valvules cardiaques. Ce test peut être conseillé aux personnes atteintes d'athérosclérose ou d'insuffisance coronarienne. On y a recours pour évaluer le flux sanguin dans les artères coronariennes (vaisseaux sanguins qui alimentent le cœur), l'artère carotide (principale artère du cou), les principales artères des bras et des jambes, ou même à l'intérieur du cœur en soi (échocardiogramme).

2.6.5 Angiographie coronarienne (coronarographie)

Une angiographie coronarienne (aussi appelée coronarographie) est un test qui consiste à prendre des radiographies des artères coronariennes et des vaisseaux qui alimentent le cœur. Au cours de cette intervention, une coloration spéciale, soit un produit iodé, est injectée dans les artères coronariennes à partir d'un cathéter (tube long et étroit) inséré dans un vaisseau sanguin ; chacun d'entre eux devient alors visible à la radiographie. L'angiographie permet aux médecins d'observer la circulation sanguine dans le cœur, de façon distincte, et parfois même de cerner d'éventuels problèmes au niveau des artères coronaires.

2.6.6 Enregistrement Holter ou surveillance d'événements

L'enregistrement ECG par la méthode Holter (appelée aussi enregistrement Holter) est habituellement utilisé pour diagnostiquer les anomalies du rythme cardiaque, plus spécifiquement pour trouver l'origine des palpitations ou des

étourdissements. Pour ce faire, vous devez porter un petit appareil d'enregistrement de l'ECG, appelé moniteur Holter, qui est relié à de petits disques métalliques (électrodes) placés sur votre poitrine, qui permettent une lecture de votre pouls et de votre rythme cardiaque au cours d'une période minimale de 24 heures. Votre rythme cardiaque est alors transmis et enregistré sur bande, puis numérisé par l'ordinateur, afin d'être analysé pour découvrir ce qui cause votre arythmie. Il est à noter qu'avec certains moniteurs, vous devez appuyer sur un bouton pour enregistrer le rythme cardiaque au moment où vous ressentez des symptômes.

3. Analyse de Données Médicales pour l'entrainement des modèles de classification

Dans le domaine médical, l'utilisation de bases de données dédiées aux maladies cardiaques revêt une importance cruciale. Ces bases de données regroupent une multitude d'informations telles que des images, des signaux physiologiques et des données spécifiques liées aux affections cardiaques. Elles fournissent une mine de données essentielles sur les patients, leurs antécédents médicaux, les résultats des tests, les diagnostics et les traitements liés aux maladies cardiaques. En centralisant ces données diverses, les bases de données facilitent la recherche, l'analyse et le développement de stratégies de diagnostic et de traitement plus efficaces pour ces affections critiques.

3.1. Les Différentes Modalités de Données Utilisées dans les Maladies Cardiaques

3.1.1 Base de Données Utilisant des Signaux

- **Base de données ECG diagnostique PTB :** Il s'agit d'une collection de 549 ECG haute résolution à 15 dérivations (12 dérivations standard ainsi que les dérivations Frank XYZ), comprenant des résumés cliniques pour chaque enregistrement. De un à cinq enregistrements ECG sont disponibles pour chacun des 294 sujets, qui comprennent des sujets sains ainsi que des patients souffrant de diverses maladies cardiaques [7].

3.1.2 Base de Données Utilisant des Images

- **MESA (Multi-Ethnic Study of Atherosclerosis)** : Il s'agit d'une étude de population cardiovasculaire à grande échelle (> 6 500 participants). Son objectif est d'étudier la manifestation des maladies cardiovasculaires subcliniques à cliniques avant que les signes et symptômes ne se développent. Étant donné que les participants à MESA ont été déterminés comme étant en bonne santé lors du recrutement, les patients de ces données sont identifiés comme des sujets asymptomatiques. Les données démographiques globales de l'étude MESA comprennent des hommes et des femmes âgés de 45 à 84 ans. Une partie des données MESA comprend des IRM (2 450 cas) de l'examen de base. De plus, environ 300 modèles ventriculaires gauches 3D ont été dérivés de manière semi-automatique en utilisant le logiciel CIM [8].

3.1.3 Base de Données Utilisant des Données Démographiques et Cliniques

- **Framingham heart study (FHS)** : La base de données de l'étude sur le cœur de Framingham comprend comprend plus de 4 240 enregistrements, répartis sur 16 colonnes et comprenant 15 attributs. L'objectif de cette base de données est de prédire si le patient présente un risque de maladie coronarienne future (CHD) sur 10 ans [9].
- **Cleveland Clinic Heart Disease Dataset :** Cette base de données comprend 303 observations, 13 caractéristiques et 1 attribut cible. Les 13 caractéristiques incluent les résultats des tests de diagnostic non invasifs, ainsi que d'autres informations pertinentes sur les patients. L'attribut cible comprend le résultat de l'angiographie coronarienne invasive qui représente la présence ou l'absence de maladie coronarienne chez le patient, 0 représente l'absence de maladie coronarienne (CHD), tandis que les étiquettes 1 à 4 représentent la présence de CHD. La plupart des recherches utilisant cet ensemble de données se sont concentrées simplement sur la tentative de distinguer la présence (valeurs 1, 2, 3,4) de l'absence (valeur 0) [10].

3.2 Prétraitement des données

Dans le domaine de l'analyse des données et de l'intelligence artificielle, le prétraitement des données est crucial pour obtenir des résultats précis et optimiser les performances des modèles et des applications. Les types prétraitement des données varient en fonction des besoins

spécifiques de chaque type de données, comprenant prétraitement de la qualité, le nettoyage, la sélection et la transformation. Cette section explore ces différentes méthodes et leur application pour améliorer la qualité et l'utilité des données dans divers contextes et applications [11].

3.2.1 Filtrage de valeurs manquantes

Les valeurs manquantes constituent un problème récurrent dans les ensembles de données du monde réel, car ces données sont soumises à des contraintes physiques et opérationnelles. Par exemple, si des données sont capturées par des capteurs provenant d'une source particulière, le capteur peut cesser de fonctionner pendant un certain temps, entraînant des données manquantes. De même, différents ensembles de données présentent différents problèmes qui entraînent des points de données manquants.

Nous devons gérer ces valeurs manquantes pour exploiter de manière optimale les données disponibles. Voici quelques méthodes éprouvées [11] :

- **Supprimer les échantillons avec des valeurs manquantes** : ceci est déterminant lorsque le nombre d'échantillons est élevé et que le nombre de valeurs manquantes dans une ligne/un échantillon est élevé. Ce n'est pas une solution recommandée dans les autres cas, car elle entraîne d'importantes pertes de données.
- **Remplacez les valeurs manquantes par zéro :** cette technique fonctionne parfois pour des ensembles de données de base, car les données en question supposent zéro comme nombre de base, ce qui signifie que la valeur est absente. Cependant, dans la plupart des cas, zéro peut signifier une valeur en soi. Par exemple, si un capteur génère des valeurs de température et que l'ensemble de données appartient à une région tropicale. De même, dans la plupart des cas, si les valeurs manquantes sont renseignées avec 0, cela pourrait induire le modèle en erreur. 0 ne peut être utilisé en remplacement que lorsque l'ensemble de données est indépendant de son effet. Par exemple, dans les données de facture téléphonique, une valeur manquante dans la colonne montant facturer peut être remplacée par zéro, car cela peut indiquer que l'utilisateur n'a pas souscrit au forfait ce mois-là.
- **Remplacez la valeur manquante par la moyenne, la médiane ou le mode :** vous pouvez résoudre le problème ci-dessus, résultant d'une mauvaise utilisation de 0, en utilisant des fonctions statistiques telles que la moyenne, la médiane ou le mode pour remplacer les valeurs manquantes. Même s'il s'agit également d'hypothèses, ces valeurs

ont plus de sens et constituent des approximations plus proches qu'une valeur unique telle que 0.

- **Interpoler les valeurs manquantes :** l'interpolation permet de générer des valeurs dans une plage basée sur une taille de pas donnée. Par exemple, s'il y a 9 valeurs manquantes dans une colonne entre les cellules de valeurs 0 et 10, l'interpolation remplira les cellules manquantes avec des nombres de 1 à 9. Naturellement, l'ensemble de données doit être trié selon une variable plus fiable (comme le numéro de série) avant interpolation.
- **Construisez un modèle avec d'autres fonctionnalités pour prédire les valeurs manquantes :** Ici, un algorithme étudie toutes les variables sauf la variable cible réelle (car cela entraînerait une fuite de données). La variable cible de cet algorithme devient la fonctionnalité avec des valeurs manquantes. Le modèle, s'il est bien entraîné, peut prédire les points manquants et fournir les approximations les plus proches.

3.2.2 Transformation des données

La transformation est une étape cruciale du prétraitement des données dans l'apprentissage automatique. Elle vise à modifier les données d'une forme à une autre pour les rendre plus utiles pour l'apprentissage automatique.

Les objectifs de la transformation sont multiples :

- Améliorer la performance des modèles d'apprentissage automatique.
- Faciliter la comparaison des données.
- Éviter les problèmes de sur justement et de sous-ajustement.
- Rendre les données plus compatibles avec les algorithmes d'apprentissage automatique.

❖ **Normalisation**

La normalisation est une technique de prétraitement des données dans l'apprentissage automatique qui vise à mettre à l'échelle les données afin qu'elles soient toutes dans la même plage.

- **Description :** Cette technique met à l'échelle les caractéristiques entre 0 et 1.

- **Équation :**

$$X\_normalisé = \frac{x - xmin}{xmax - xmin}$$

- ✓ x : est la valeur à normaliser.
- ✓ $xmin$: est la valeur minimale de la caractéristiques
- ✓ $xmax$: est la valeur maximale de la caractéristiques

❖ **Standardisation**

La standardisation est un processus formel qui vise à définir des normes pour la manipulation des données dans l'apprentissage automatique.

- **Description :** Cette technique met à l'échelle les caractéristiques pour avoir une moyenne de 0 et un écart-type de 1.
- **Équation :**

$$X\_normalisé = \frac{x - \mu}{\sigma}$$

4 Evaluation du Modèle

L'évaluation d'un modèle de classification est une étape cruciale pour s'assurer de sa fiabilité et de son efficacité dans le monde réel. Pour ce faire, il est essentiel de tester le modèle sur des données inédites, non utilisées lors de l'entraînement.

4.1 Matrice de confusion

La matrice de confusion est un outil précieux pour visualiser et analyser les performances du modèle face à de nouvelles situations. En décortiquant la matrice de confusion, on peut identifier les forces et les faiblesses du modèle, et ainsi le perfectionner pour une meilleure performance

La matrice de confusion résume les résultats de prédiction pour un problème de classification en comparant les données réelles pour une variable cible à celles prédites par le modèle [7]. Elle présente les prédictions du modèle, qu'elles soient correctes ou fausses, de manière claire et concise dans quatre catégories distinctes :

True Positive (TP) : la prédiction et la valeur réelle sont positives. Exemple : Une personne malade et prédite comme malade.
True Négative (TN) : la prédiction et la valeur réelle sont négatives. Exemple : Une personne saine et prédite comme saine.
False Positive (FP) : la prédiction est positive alors que la valeur réelle est négative. Exemple : Une personne saine et prédite comme malade.
False Négative (FN) : la prédiction est négative alors que la valeur réelle est négative. Exemple : Une personne malade et prédite comme saine.

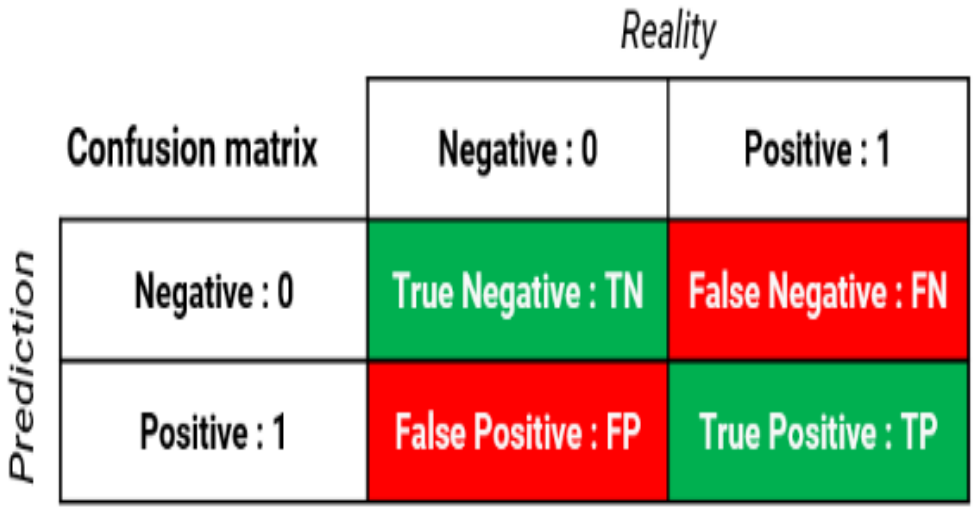

Figure 2 la matrice de confusion. [13]

4.2 Métriques d'Évaluation

Diverses métriques peuvent être dérivées à partir du tableau de confusion pour faciliter son interprétation. Parmi celles-ci figurent l'exactitude (Accuracy), la précision (Precision), le

rappel (Recall) et le score F1 (F-measure). Ces indicateurs permettent de mieux apprécier la qualité de précision du modèle [12].

4.2.1 Exactitude :

Ce paramètre correspond à la somme des vrais positifs (TP) et des vrais négatifs (TN), divisée par le nombre total d'instances. Des valeurs élevées de cette métrique sont généralement souhaitables. Elle peut également être calculée à l'aide de la formule suivante :

$$\boldsymbol{Exactitute} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + TN + FP + FN}}$$

4.2.2 Précision :

La précision indique le rapport entre les prévisions positives correctes (TP) et le nombre total de prévisions positives. La métrique de précision révèle le nombre de classes prédites qui sont correctement étiquetées.

$$\boldsymbol{Précision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FP}}$$

4.2.3 Rappel :

Mesure la capacité du modèle à prédire les classes positives réelles. Il s'agit du rapport entre les vrais positifs (TP) prédits et ce qui a été réellement étiqueté. La métrique de rappel révèle le nombre de classes prédites correctes.

$$\boldsymbol{Rappel} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FN}}$$

4.2.4 F-measure :

Le score F1 est une fonction de précision et de rappel. Il est nécessaire quand vous recherchez l'équilibre entre précision et rappel.

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

$$F - measure = \frac{2 * TP}{2 * TP + FP + FN}$$

5 Revue littérature

L'importance du lien entre l'intelligence artificielle et le domaine de la santé, en particulier dans le diagnostic des maladies cardiaques, s'accroît dans le monde moderne. L'intégration de l'IA dans la cardiologie ouvre des perspectives prometteuses pour améliorer les processus de diagnostic en permettant une détection précoce et précise des maladies cardiaques. Les systèmes d'IA sont capables d'analyser de vastes ensembles de données médicales, y compris des images médicales et des données biométriques, afin de repérer les signes précoces de maladies cardiaques, ce qui apporte un soutien décisionnel précieux aux médecins. Cette convergence entre l'intelligence artificielle et la santé cardiaque ouvre la voie à des approches médicales plus personnalisées et réactives, renforçant ainsi les efforts de prévention et de gestion des maladies cardiaques

Dans cette section, nous présentons certaines des solutions récentes, le tableau 1 résume les résultats des solutions étudiées :

Dans [14], l'étude présentée par Shadman Nashif et al, 2018, une conception provisoire d'un système de prédiction des maladies cardiaques basé sur le Cloud a été proposée pour détecter les maladies cardiaques imminentes à l'aide de techniques d'apprentissage automatique. Pour une détection précise des maladies cardiaques, une technique d'apprentissage automatique efficace devrait être utilisée, issue d'une analyse distinctive parmi plusieurs algorithmes d'apprentissage automatique dans une plateforme de fouille de données open source basée sur Java, WEKA. L'algorithme proposé a été validé à l'aide de deux bases de données open source largement utilisées, où une validation croisée à 10 volets est appliquée afin d'analyser les performances de la détection des maladies cardiaques. Un niveau de précision de 97,53% a été obtenu avec l'algorithme SVM, ainsi qu'une sensibilité et une spécificité de 97,50% et 94,94% respectivement.

Dans [15], LIQAT ALI et al, 2019 ont conçu un système intelligent pour prédire les maladies cardiaques en utilisant deux modèles. Le premier modèle SVM est linéaire et régularisé L1. Ce qui lui permet d'éliminer les caractéristiques non pertinentes en réduisant leurs coefficients à zéro. Le deuxième modèle SVM est régularisé L2 et est utilisé comme modèle prédictif. Pour optimiser les deux modèles, ils ont proposé un algorithme de recherche par grille hybride (HGSA) capable d'optimiser les deux modèles simultanément. L'efficacité de la méthode proposée est évaluée à l'aide de six métriques d'évaluation différentes : précision, sensibilité, spécificité, coefficient de corrélation de Matthews (MCC), courbes ROC et aire sous la courbe (AUC). Les résultats expérimentaux confirment que la méthode proposée améliore les performances d'un modèle SVM classique de 3.3 %.

Dans [16], l'étude proposée par Gupta et al., 2022, diverses techniques d'apprentissage supervisé, telles que le k-plus proche voisin, l'arbre de décision, la régression logistique, le naïf bayésien et le modèle de machine à vecteurs de support (SVM), sont utilisées pour prédire les maladies cardiaques à l'aide d'un ensemble de données collecté dans le référentiel de l'Université de Californie à Irvine (UCI). Les résultats montrent que la régression logistique était meilleure que tous les autres classificateurs supervisés en termes de métriques de performance. Le modèle est également moins risqué, car le nombre de faux négatifs est faible par rapport aux autres modèles, comme le montre la matrice de confusion de tous les modèles. De plus, des techniques d'ensemble peuvent être utilisées pour améliorer la précision du classificateur.

Dans le travail proposé par Chiradeep Gupta et al, 2022 [17], une variété de technique de machine Learning supervisé telles que K-Nearest Neighbors, l'arbre de décision, la régression logistique, Naïve Bayes et le modèle de machine à vecteurs de support (SVM) sont utilisées pour prédire les maladies cardiaques à l'aide d'un ensemble de données collectées auprès du référentiel de l'Université de Californie, Irvine (UCI).En utilisant "70% des données ont été utilisées pour l'entraînement et 30% pour l'évaluation

Dans [18], A Angel Nancy, 2022 s'est efforcée d'évaluer les modèles de prédiction séquentielle sur l'ensemble de données des maladies cardiaques avec des modèles d'apprentissage en profondeur comprenant le modèle LSTM générique et le FIS combiné avec LSTM (FLSTM) aux côtés du modèle proposé. Le système exploite les ensembles de données sur les maladies cardiaques de Cleveland et de Hongrie, accessibles depuis le référentiel en ligne de l'apprentissage automatique et de l'exploration de données de l'Université de Californie, Irvine (UCI). Les ensembles de données originaux sur les maladies cardiaques de Cleveland et de Hongrie comprennent respectivement 303 et 294

enregistrements, avec 14 caractéristiques. Ces enregistrements ont été augmentés à 100 000 enregistrements en utilisant Mockaroo, l'outil de génération d'ensembles de données, pour vérifier la robustesse du modèle d'apprentissage en profondeur proposé. Ainsi, le système est validé en utilisant 100 000 enregistrements segmentés en 70 % pour les tâches d'entraînement et 30 % pour les tâches de test.

Dans [19] [A Angel Nancy et al, 2023], les auteurs ont proposé un système de santé intelligent assisté par le fog pour diagnostiquer les maladies cardiaques ou cardiovasculaires. Il a combiné un système d'inférence floue (FIS) avec la variante du modèle de réseau neuronal récurrent de l'unité récurrente à porte (GRU) pour les tâches de prétraitement et d'analyse prédictive. Le système proposé présente des résultats de performance considérablement améliorés, avec une précision de classification de 99.125%.

Référence	**Dataset**	**Année**	**Algorithmes**	**Métrique d'évaluation**
[14]	UCI Cleveland	2018	**SVM** MLP Simple logistic NB RF	**SVM** Accurency = 97.53% Sensitivity = 97.5% Specificity = 94.94%
[15]	UCI Cleveland	2019	**SVM**	**SVM** Accurency =92.22%
[16]	UCI Cleveland	2021	**LR** SVM NB DT KNN RF	**LR** Accuracy= 92.3%
[17]	UCI Cleveland	2022	**LR** SVM NB DT KNN	**LR** Sensitivity = 96.08% Specificity = 87.5% Precision = 90.74%

			RF	F1 Score = 93.34%
[18]	UCI Cleveland	2022	**Bi-LSTM**	**Bi-LSTM** Accuracy= 98.86% Precision= 98.9% Sensitivity= 98.8% Specificity=98.89% F-measure=98.86%
[19]	UCI Cleveland	2023	**GRU**	**GRU** Accuracy =99.125%

Tableau 1 Comparaison des travaux étudiés

6. Analyse critiques

L'analyse des travaux étudiés met en lumière l'importance de l'utilisation des techniques d'apprentissage automatique dans la prédiction et la détection des maladies cardiaques. Plusieurs approches sont discutées, dont l'utilisation d'algorithmes d'apprentissage automatique tels que les SVM (Support Vector Machines), la régression logistique, ainsi que d'autres techniques supervisées pour développer des modèles prédictifs de maladies cardiaques. L'accent est particulièrement mis sur l'importance de la précision dans la détection de ces maladies, soulignant la nécessité d'évaluer les performances des modèles à l'aide de différentes métriques pour garantir leur fiabilité et leur efficacité. En résumé, ces paragraphes illustrent comment les avancées en matière d'apprentissage automatique peuvent être appliquées de manière efficace pour améliorer la détection et la prédiction des maladies cardiaques, ce qui pourrait avoir un impact significatif sur les soins de santé préventifs et l'intervention médicale précoce.

La répartition 70% train et 30% test présente plusieurs inconvénients, principalement liés à la réduction de la quantité de données disponibles pour l'entraînement du modèle. Cela peut limiter la capacité du modèle à apprendre et à généraliser efficacement, augmenter la variance des estimations des paramètres et réduire les performances globales, en particulier pour les

modèles plus complexes qui nécessitent des volumes de données plus importants pour un apprentissage efficace. En conséquence, bien que cette répartition puisse offrir un avantage en termes d'évaluation plus robuste, elle peut nuire à l'efficacité de l'apprentissage, surtout lorsque les données disponibles sont limitées.

Dans le domaine de l'analyse des données et de la prédiction à l'aide des techniques d'intelligence artificielle, l'attention est dirigée vers le choix de l'approche la plus adaptée en fonction de la taille de la base de données disponible. Lorsqu'il s'agit de traiter de petites bases de données telles que la base de données UCI Cleveland, l'utilisation de techniques d'apprentissage automatique traditionnelles est plus efficace que l'apprentissage profond. Cela s'explique par le fait que des algorithmes d'apprentissage automatique tels que l'arbre de décision, la forêt aléatoire et la machine à vecteurs de support peuvent fournir de bons résultats et analyser les modèles efficacement sans avoir besoin de grandes quantités de données. En revanche, les modèles d'apprentissage profond, tels que les réseaux de neurones profonds (DNN) et les réseaux de neurones convolutionnels (CNN), dépendent de grandes quantités de données pour entraîner leurs modèles complexes efficacement. Par conséquent, lors de l'utilisation de la base de données UCI Cleveland, qui contient des données relativement limitées, les techniques d'apprentissage automatique sont plus appropriées car elles réduisent les risques de sous-apprentissage et offrent de meilleures performances en termes de prédiction et d'analyse.

7. Comparaison des méthodes et des résultats

Chaque étude apporte des contributions uniques à la prédiction des maladies cardiaques, avec des différences dans les méthodes, les ensembles de données et les résultats de performance. Les techniques avancées telles que l'apprentissage profond, y compris les modèles LSTM et GRU, ont montré des performances exceptionnelles en termes de précision. Ces techniques sont capables de traiter des données complexes et d'améliorer considérablement la précision des prédictions, ce qui reflète leur puissance dans ce domaine. D'autre part, les méthodes traditionnelles comme les SVM et la régression logistique restent robustes et efficaces, car elles peuvent obtenir des résultats proches de ceux des modèles d'apprentissage profond plus complexes, avec un taux faible de faux négatifs, ce qui en fait une option valable dans certains cas.

Les approches d'amélioration, telles que l'utilisation de différentes algorithmes de recherche et d'optimisation, ont montré une amélioration des performances, renforçant ainsi l'efficacité des modèles dans la prédiction précise des maladies cardiaques. De plus, l'utilisation de diverses métriques d'évaluation enrichit la comparaison et la compréhension des modèles prédictifs de manière globale. Ces métriques peuvent inclure la précision, la sensibilité, la spécificité, le coefficient de corrélation de Matthews, les courbes ROC et l'aire sous la courbe, fournissant ainsi une vue d'ensemble des performances des modèles et aidant à identifier les points forts et faibles de chaque approche.

En général, chaque approche présente des avantages spécifiques pouvant être exploités en fonction du contexte et des données disponibles. Les techniques avancées offrent une plus grande précision, tandis que les méthodes traditionnelles apportent stabilité et facilité d'application. Les améliorations et les multiples métriques d'évaluation ajoutent une valeur significative au processus de prédiction des maladies cardiaques, aidant ainsi à choisir l'approche la plus appropriée pour obtenir les meilleurs résultats.

8.Conclusion

Le chapitre examine en détail l'importance de l'utilisation des technologies de l'intelligence artificielle et de l'apprentissage automatique dans la prédiction et la gestion des maladies cardiaques. Il commence par définir les maladies cardiaques, en mettant en lumière leurs différents types, symptômes et facteurs de risque. Ensuite, il explore les concepts clés de l'apprentissage automatique, de l'intelligence artificielle et de l'apprentissage profond, ainsi que leur application spécifique dans le domaine de la cardiologie. Le chapitre examine également les différents types de jeux de données utilisés dans la recherche sur les maladies cardiaques, ainsi que les méthodes de prétraitement nécessaires pour assurer la qualité des données. Enfin, il passe en revue les études antérieures dans ce domaine, mettant en évidence les progrès réalisés et les défis à relever pour une utilisation plus efficace des techniques d'intelligence artificielle dans la lutte contre les maladies cardiaques.

Chapitre 2 :

Analyse, prétraitement et entraînement du modèle de classification

1. Introduction

Dans ce chapitre, nous allons explorer en détail notre travail, en commençant par une architecture générale qui offre une vue d'ensemble claire de notre système. Ensuite, nous plongerons dans l'analyse approfondie du dataset UCI Cleveland, en fournissant une description détaillée de ses caractéristiques et de ses attributs. Nous discuterons également du prétraitement nécessaire pour garantir la qualité des données, en mettant particulièrement l'accent sur la standardisation pour assurer une comparabilité efficace entre les variables. Ensuite, nous aborderons la division du dataset en ensembles d'entraînement (80%) et de test (20%), une étape cruciale pour évaluer la performance de nos modèles. Enfin, nous appliquerons plusieurs algorithmes de machine learning, notamment SVM, KNN, Random Forest et Arbres de Décision, pour explorer et comparer leur efficacité dans la prédiction des résultats.

2. Architecture générale

Le schéma dans figure 3 illustre le processus d'analyse des données et d'application de l'apprentissage automatique avec l'utilisation d'un système en temps réel. Il commence par la collecte des données et leur analyse, suivies d'un prétraitement consistant à supprimer les valeurs aberrantes et à standardiser les données. Ensuite, les données sont divisées en un ensemble d'entraînement (80%) et un ensemble de test (20%). Les algorithmes d'apprentissage supervisé (comme SVM, DT, KNN, RF) sont appliqués et le meilleur modèle est sélectionné (SVM). Un algorithme génétique est utilisé pour sélectionner les caractéristiques les plus influentes, puis le modèle est appliqué à de nouvelles données. Enfin, le système en temps réel est utilisé pour prendre des décisions immédiates, mettre à jour le modèle, envoyer des alertes et notifications, et effectuer une surveillance continue pour garantir des performances optimales.

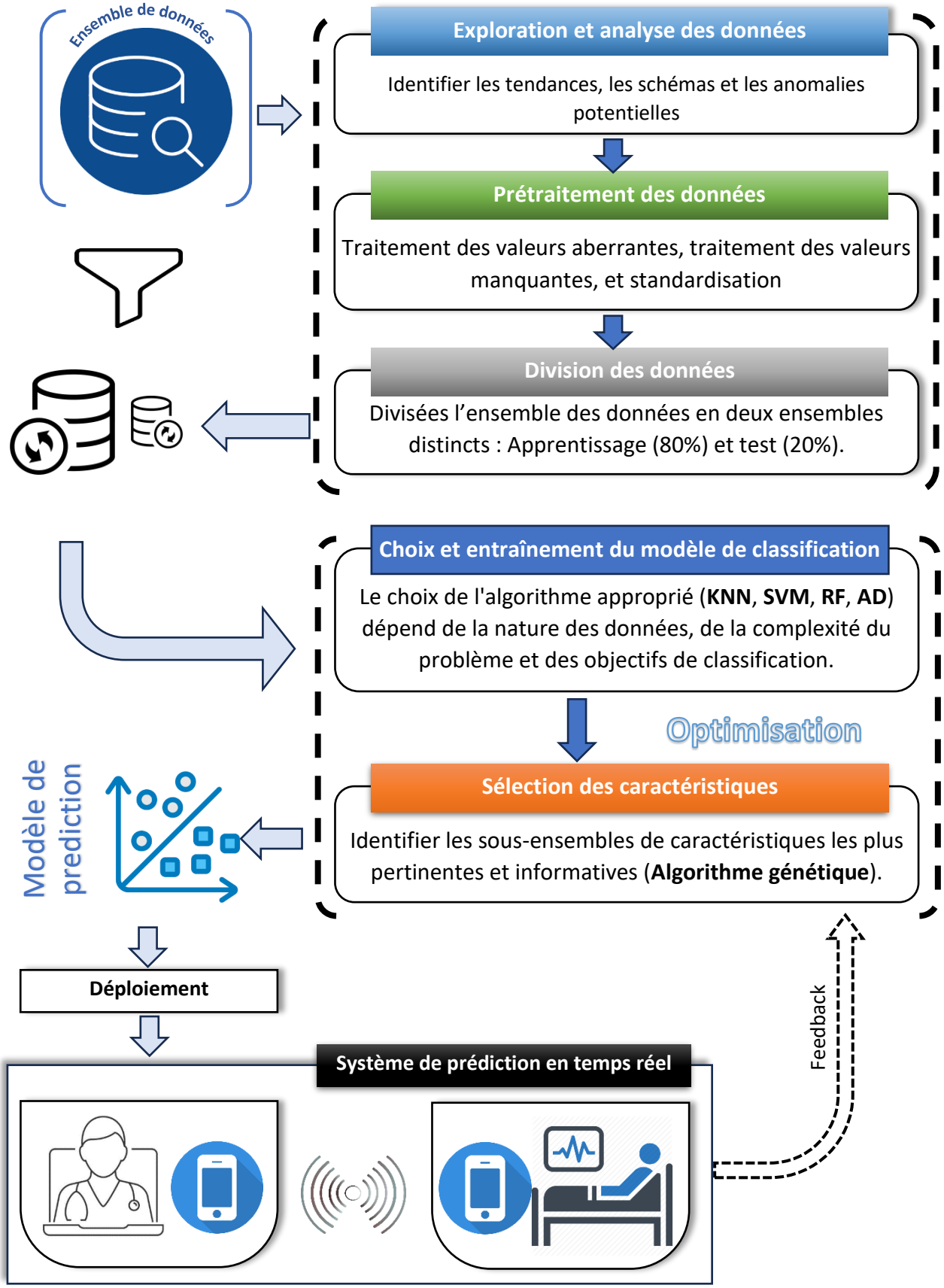

Figure 3 architecture générale de notre système

3. Le Dataset Cleveland Heart Disease:

- **Aperçu du dataset**

Le dataset utilisé est le Cleveland Heart Disease extrait du référentiel UCI. Cet ensemble est largement utilisé pour la recherche et les projets liés à la classification des maladies cardiaques. Il comprend 303 enregistrements individuels et 14 caractéristiques, extraites d'un ensemble initial de 75. La tâche de classification consiste à prédire si un individu souffre ou non d'une maladie cardiaque. (0 : Absence, 1 : présence). [20]

- **Caractéristiques du dataset**

Ce dataset contient 13 attributs et une variable cible. Il comporte 8 valeurs nominales et 5 valeurs numériques. La description détaillée de toutes ces caractéristiques est illustrée dans le tableau ci-dessous :

Attributs	**Indications**
Age	Âge des patients en années
Sex	Genre (Homme : 1 ; Femme : 0)
Cp	Type de douleur thoracique ressentie par le patient. Ce terme est classé en 4 catégories : – 0 : Angine typique – 1 : Angine atypique – 2 : Douleurs non angineuses – 3 : Asymptomatiques
Trestbps	Niveau de tension artérielle au repos en mm/HG
Chol	Cholestérol sérique en mg/dl
Fbs	la glycémie à jeun > 120 mg/dl représente 1 en cas de vrai et 0 en cas de faux

Restecg	Les résultats de l'électrocardiogramme au repos sont représentés en 3 valeurs distinctes : – 0 : Normal – 1 : Présentant une anomalie de l'onde ST-T (inversions de l'onde T et/ou élévation ou dépression ST >0,05 mV) – 2 Montrant une hypertrophie ventriculaire gauche probable ou certaine par Critères d'Estes
Thalach	Fréquence cardiaque maximale atteinte
Exang	Angine induite par l'exercice (0 = NON, 1 = Oui)
Oldpeak	Dépression du segment ST induite par l'exercice par rapport à l'état de repos
Slope	Pente du segment ST maximal pendant l'exercice : – 0 : Montée ; – 1 : Plate ; – 2 : Descente
Ca	Nombre de gros vaisseaux colorés par fluoroscopie (0-3)
Thal	Un trouble sanguin appelé Thalassémie Thalassémie (3 = normal, 6 = défaut fixe, 7 = défaut réversible)
Num	Diagnostic de la maladie cardiaque – 1 signifie que le patient souffre d'une maladie cardiaque (présence) – 0 signifie que le patient est normal (absence).

Tableau 2 Caractéristiques du dataset Cleveland et leurs descriptions

4. Analyse des données

4.1. Analyse de la forme

- **La taille** : Il se compose de 303 lignes et 14 colonnes il y a 13 caractéristiques avec le 'num' le résultat

- **Les colonnes** : Les colonnes de l'ensemble de données sont : ['age', 'sex', 'cp','trestbps', 'chol', 'fbs', 'restecg', 'thalach','exang', 'oldpeak', 'slope', 'ca', 'thal','num'].
- **Le type** : L'ensemble de données et de type : Object.

```
Shape of DataFrame: (303, 14)
age          67.0
sex           1.0
cp            4.0
trestbps    160.0
chol        286.0
fbs           0.0
restecg       2.0
thalach     108.0
exang         1.0
oldpeak       1.5
slope         2.0
ca            3.0
thal          3.0
num             1
Name: 1, dtype: object
Index(['age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach',
       'exang', 'oldpeak', 'slope', 'ca', 'thal', 'num'],
      dtype='object')
```

Figure 4 Visualisation de la form

- **Statistiques :**

Dans cette section, nous utilisons la méthode df.describe() pour générer des statistiques descriptives telles que la moyenne, la médiane, l'écart-type, etc., pour chaque variable. Cette approche nous permet de mieux comprendre la distribution des données et de détecter d'éventuelles valeurs aberrantes.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.438944	0.679868	3.158416	131.689769	246.693069	0.148515	0.990099	149.607261	0.326733	1.039604
std	9.038662	0.467299	0.960126	17.599748	51.776918	0.356198	0.994971	22.875003	0.469794	1.161075
min	29.000000	0.000000	1.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000
25%	48.000000	0.000000	3.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000
50%	56.000000	1.000000	3.000000	130.000000	241.000000	0.000000	1.000000	153.000000	0.000000	0.800000
75%	61.000000	1.000000	4.000000	140.000000	275.000000	0.000000	2.000000	166.000000	1.000000	1.600000
max	77.000000	1.000000	4.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000

Figure 5 Description de l'ensemble de données.

Par exemple, d'après les données fournies par la figure 5 :

- **Pour l'âge :** la moyenne est de 54.44 ans avec un écart type de 9.04. L'âge minimum est de 29 ans, le premier quartile est à 48 ans, la médiane à 56 ans, le troisième quartile à 61 ans, et l'âge maximum est de 77 ans.
- **Pour la pression artérielle au repos (Testbps) :** la moyenne est de 131.69 mmHg avec un écart type de 17.60. La pression artérielle minimale est de 94 mmHg, le premier quartile est à 120 mmHg, la médiane à 130 mmHg, le troisième quartile à 140 mmHg,et la pression artérielle maximale est de 200 mmHg

4.2.Analyse uni-variée

L'objectif de l'analyse uni-variée est de décrire et d'évaluer comment les valeurs d'une variable se répartissent dans l'ensemble des données.

4.2.1. Visualisation de la classe résultat

Le graphe de l'attribut "num", qui est l'étiquette ou la variable cible, est important pour évaluer l'équilibre des données, un aspect essentiel pour obtenir des résultats précis. Cet attribut joue un rôle crucial en indiquant l'absence ou la présence de la maladie cardiaque chez les patients. Dans la figure 6, le graphe représente la répartition de la variable "num". Nous observons 138 cas représentant 45,54 % des personnes atteintes de maladie cardiaque, tandis que 165 cas représentent 54,45 % des personnes non atteintes de maladie cardiaque. Cette répartition quasi équilibrée entre les personnes malades et non malades indique que les données sont presque équilibrées.

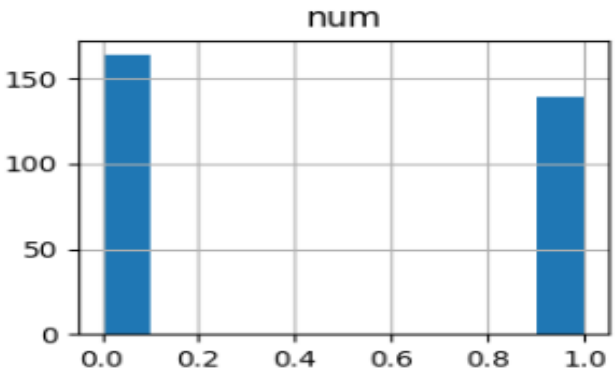

Figure 6 Graphe de l'attribut « num«

1.1.1. Visualisation des caractéristiques qualitative

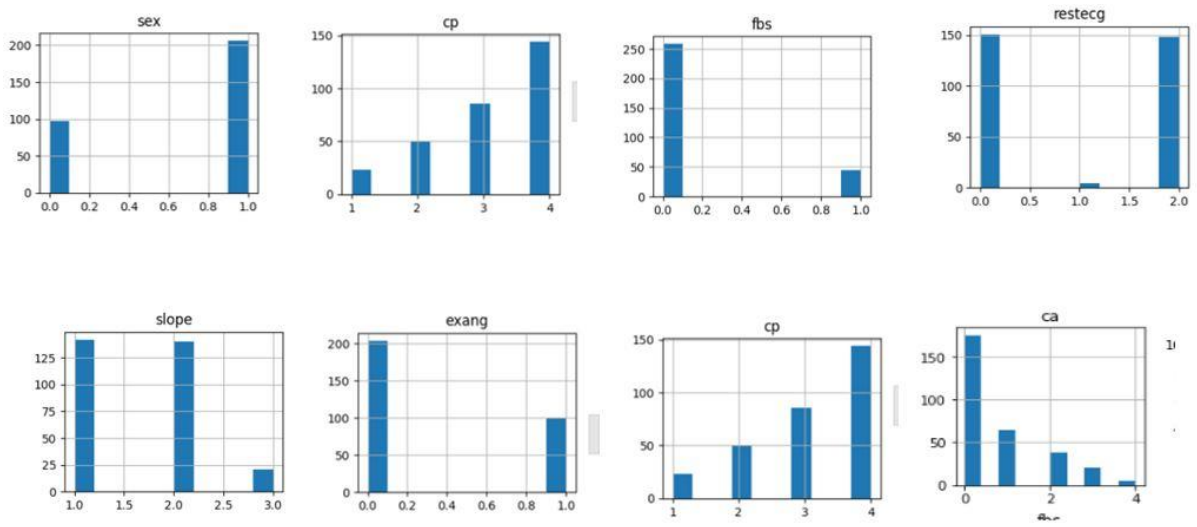

Figure 7 les graphes des caractéristiques qualitatifs

Les graphiques à barres fournissent une vue d'ensemble claire des distributions des différentes variables dans l'ensemble de données étudié. Les principales observations sont les suivantes :

- **Sex :** Le graphique montre deux barres, une pour 0 (femmes) et une plus grande pour 1 (hommes), indiquant qu'il y a plus d'hommes que de femmes dans cet ensemble de données.
- **Cp** : Graphique à 4 barres représentant les 4 types de douleurs thoraciques catégorisées de 0 à 3 classés du plus grand au plus petit (asymptomatique douleur non angineuse angine atypique, angine typique)
- **Fbs :** Graphique à barres binaire avec une barre plus petite pour 0 (glycémie à jeun <= 120mg/dl) et une barre plus grande pour 1 (glycémie > 120mg/dl).
- **Restecg :** Graphique à 3 barres représentant les 3 catégories de résultats d'électrocardiogramme au repos allant de 0 à 2.
- **Exang :** Graphique binaire avec une barre plus petite pour 0 (absence d'angine induite par l'exercice) et une plus grande pour 1 (présence).

Slope : Graphique à 2 barres pour les 2 catégories de pente du segment ST d'exercice max (0 et 1).

Ces résultats montrent que l'ensemble de données présente une diversité notable dans plusieurs variables clés, avec une légère prédominance de certaines caractéristiques. Cela suggère que l'échantillon est relativement équilibré en termes de certaines mesures, ce qui est crucial pour des modèles de machine learning robustes et pour tirer des conclusions fiables. La répartition équilibrée de plusieurs variables clés indique également que les données sont appropriées pour une analyse approfondie des relations entre ces variables et les résultats de santé, tels que la présence de maladies cardiaques.

1.1.2. Visualisation des caractéristiques quantitative

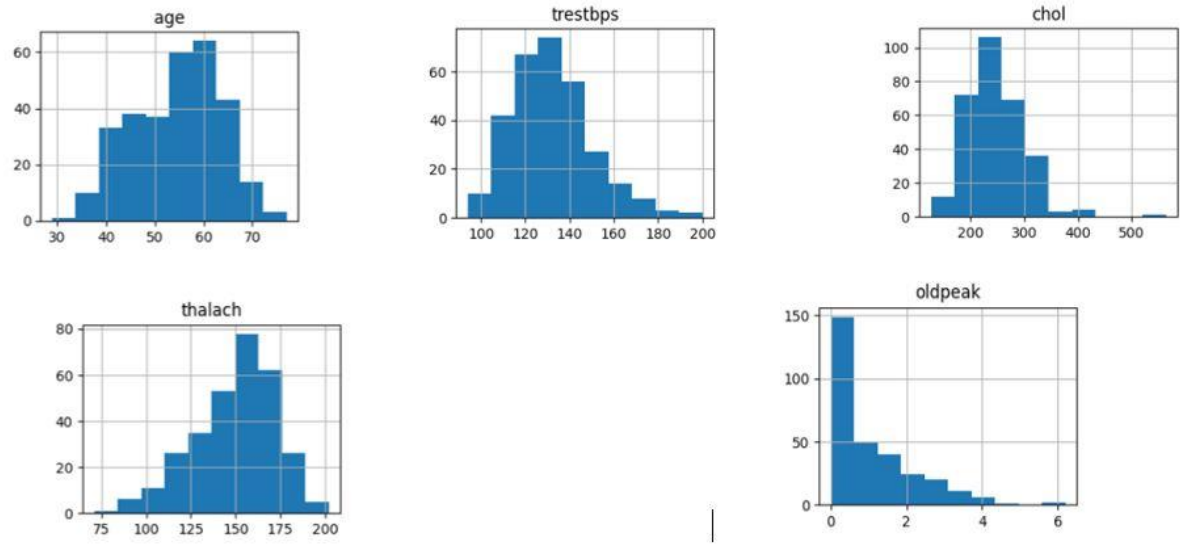

Figure 8 les graphes des caractéristiques quantitatives.

Suite aux informations fournies sur les variables. Nous pouvons retirer les observations suivantes :

- **Age :** La distribution des âges est approximativement normale, avec une concentration plus élevée autour de la tranche d'âge de 52 à 62 ans. Cela indique que la majorité es individus de cet ensemble de données se trouvent dans cette tranche d'âge. Il y a moins d'individus dans les tranches d'âge plus jeunes (moins de 40 ans) et plus âgées (plus de 70 ans).
- **trestbps** : La distribution des valeurs de la tension artérielle au repos (trestbps) dans notre dataset varie de 100 à 200 mm Hg. La majorité des valeurs se situent entre 110 et 150 mm Hg, avec un pic autour de 130 à 140 mm Hg, indiquant une tendance vers une

tension artérielle normale à légèrement élevée. Les valeurs au-dessus de 160 mm Hg sont rares, suggérant que l'hypertension sévère est peu fréquente. Cette distribution normale, avec peu de valeurs extrêmes, est bénéfique pour les modèles d'apprentissage automatique, facilitant des prédictions plus robustes et précises des issues cardiovasculaires.

- **Chol :** La distribution des niveaux de cholestérol (chol) dans notre ensemble de données varie de 100 à 500 mg/dL. La majorité des valeurs se situent entre 150 et 350 mg/dL, avec un pic notable autour de 240 mg/dL. Cela indique que la plupart des individus ont des niveaux de cholestérol normaux à légèrement élevés. Les valeurs dépassant 300 mg/dL sont rares, suggérant que des niveaux de cholestérol très élevés sont peu fréquents. Ce type de distribution est utile pour les modèles d'apprentissage automatique, car il facilite l'analyse des risques associés aux différentes concentrations de cholestérol.
- Thalach : la distribution des valeurs de la fréquence cardiaque maximale (thalach) dans notre ensemble de données varie de 75 à 200 battements par minute (bpm). La majorité des valeurs se situe entre 130 et 175 bpm, avec un pic autour de 160 bpm, indiquant que cette fréquence est la plus courante. Les valeurs très basses ou très élevées sont rares. Cette distribution quasi-normale facilite l'entraînement des modèles d'apprentissage automatique, permettant des prédictions plus précises concernant la santé cardiovasculaire.
- **Oldpeak :** La distribution de la variable "oldpeak" est asymétrique à droite, avec une concentration majeure de valeurs autour de 0. La plupart des données se situent entre 0 et 2, avec quelques valeurs extrêmes allant jusqu'à 6. Cela suggère que la majorité des cas dans l'échantillon ont une faible valeur d'oldpeak, tandis que des valeurs élevées sont rares. Une analyse plus approfondie pourrait aider à comprendre les implications cliniques et à vérifier la présence d'éventuelles erreurs de mesure.

1.2. Analyse bri-variée

1.2.1. Visualisation de la fréquence des maladies cardiaques selon âge

La figure 9 représente la fréquence des patients cardiaques et non cardiaques en fonction de l'âge. Nous remarquons que pour les personnes âgées de 29 à 54 ans, le nombre de personnes non malades est supérieur au nombre de personnes malades, indiquant ainsi une fréquence plus élevée des non malades. En revanche, pour les personnes âgées de 55 à 77 ans,

la fréquence des malades est supérieure à celle des non malades. Cela révèle une différence notable : chez les moins de 54 ans, le nombre de personnes atteintes de maladies cardiaques est plus élevé, tandis que chez les plus de 55 ans, ce sont les maladies cardiaques qui sont plus fréquentes.

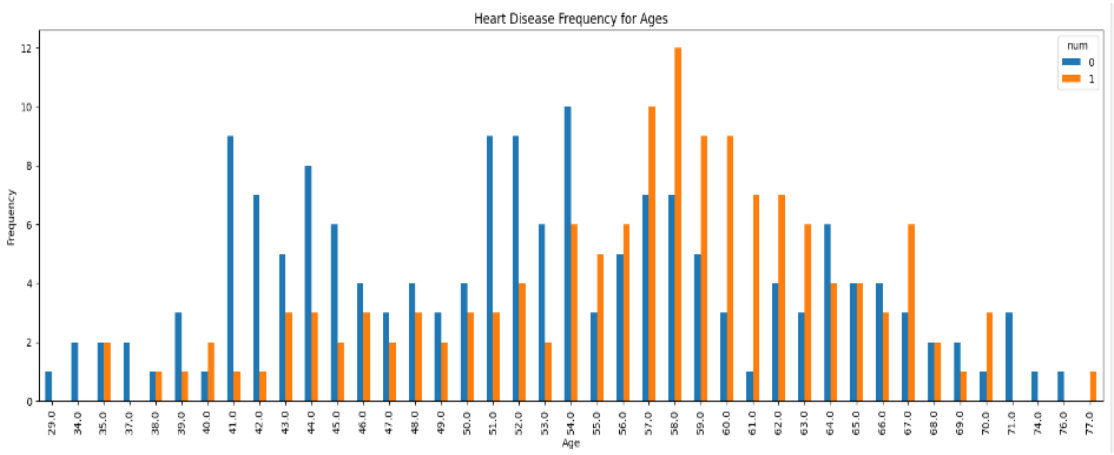

Figure 9 fréquence des maladies cardiaques selon âge.

1.2.2. Visalisation de la relation entre thalach et age

La Figure 10 montre la relation entre l'âge et la fréquence cardiaque maximale (thalach). On observe une tendance générale à la diminution de la fréquence cardiaque maximale avec l'âge. Les individus plus jeunes (moins de 40 ans) ont souvent des fréquences cardiaques maximales plus élevées, tandis que les individus plus âgés (plus de 60 ans) présentent des fréquences plus basses. Bien qu'il y ait une variabilité notable au sein de chaque groupe d'âge, cette analyse confirme les tendances attendues et souligne l'importance de l'âge dans l'évaluation de la santé cardiovasculaire, aidant ainsi à personnaliser les recommandations cliniques.

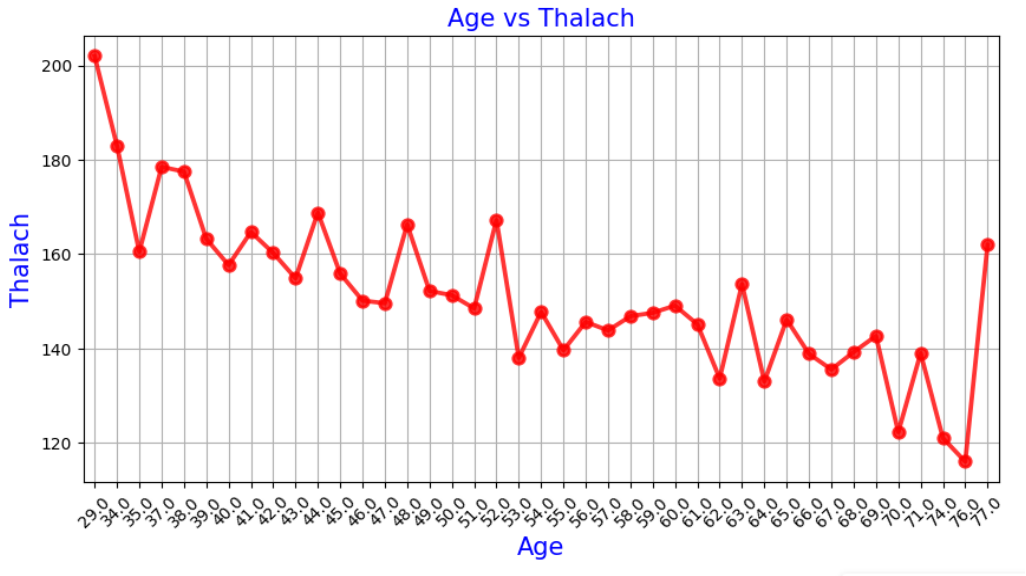

Figure 10 Visualisation de la relation entre thalach et age.

1.2.3. Corrélation entre les variables

La corrélation est une mesure statistique qui exprime la notion de liaison linéaire entre deux variables (ce qui veut dire qu'elles évoluent ensemble à une vitesse constante). C'est un outil courant permettant de décrire des relations simples sans s'occuper de la cause et de l'effet. On décrit les corrélations à l'aide d'une mesure sans unité appelée coefficient de corrélation compris entre -1 et +1 et noté *r*. [21]

À mesure que la valeur de *r* se rapproche de zéro, la relation linéaire devient plus faible. Les valeurs positives de *r* signalent une corrélation positive lorsque les valeurs des deux variables augmentent ensemble, tandis que les valeurs négatives de *r* indiquent une corrélation négative lorsque l'une des variables augmente tandis que l'autre diminue.

À travers la matrice de corrélation illustrée dans la figure 11, nous constatons qu'il existe une forte corrélation positive entre la thalassémie (thal), le nombre de vaisseaux colorés (ca), et le type de douleur thoracique (cp) avec la présence de la maladie (num). En revanche, l'âge est négativement corrélé avec la fréquence cardiaque maximale atteinte (thalach). Ces relations suggèrent que certains facteurs, comme la douleur thoracique et la thalassémie, sont fortement associés à la maladie, tandis que la fréquence cardiaque diminue avec l'âge. Ces informations peuvent guider le diagnostic et la prévention des maladies cardiovasculaires.

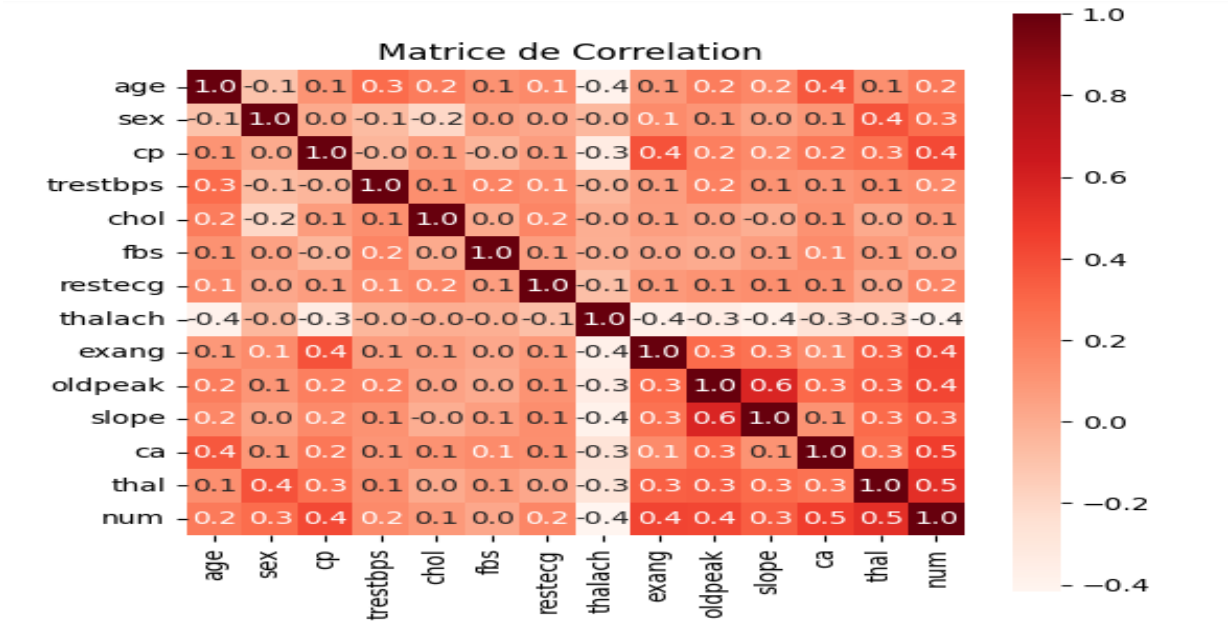

Figure 11 Matrice de corrélation.

5. Préparation des données

L'objectif de cette étape est de préparer les données de manière à les rendre compatibles avec les systèmes d'apprentissage automatique afin d'améliorer les performances de leurs modèles. Cela peut être accompli de la manière suivante :

5.1. Filtrage des valeurs manquantes :

Le processus de filtrage des valeurs manquantes (NaN) implique soit de les remplacer par d'autres valeurs telles que la moyenne de la série, le médian de la série, la moyenne des voisins, etc., soit de les supprimer complètement tout en veillant à éviter les doublons.

La figure 12 est une visualisation des valeurs manquantes sous forme de heatmap, où les cellules blanches représentent les valeurs manquantes dans notre ensemble de données. Nous remarquons que les valeurs manquantes se trouvent au niveau des caractéristiques Ca et Thal.

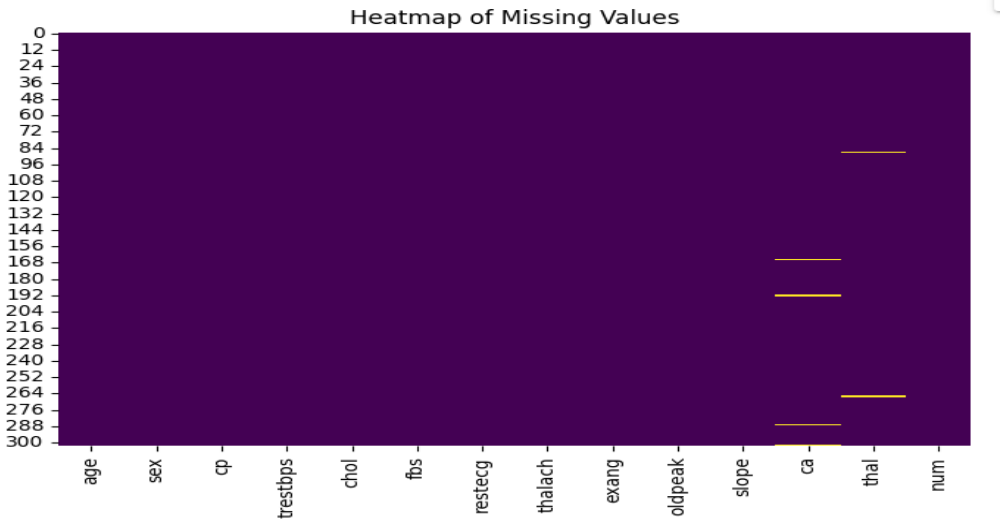

Figure 12 Visualisation des valeurs manquantes.

Pour une analyse plus approfondie, nous avons examiné les lignes qui contiennent au moins une valeur nulle. La Figure 13 indique qu'il y a six lignes dans notre base de données qui contiennent des valeurs manquantes.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	\
87	53.0	0.0	3.0	128.0	216.0	0.0	2.0	115.0	0.0	0.0	
166	52.0	1.0	3.0	138.0	223.0	0.0	0.0	169.0	0.0	0.0	
192	43.0	1.0	4.0	132.0	247.0	1.0	2.0	143.0	1.0	0.1	
266	52.0	1.0	4.0	128.0	204.0	1.0	0.0	156.0	1.0	1.0	
287	58.0	1.0	2.0	125.0	220.0	0.0	0.0	144.0	0.0	0.4	
302	38.0	1.0	3.0	138.0	175.0	0.0	0.0	173.0	0.0	0.0	

	slope	ca	thal	num
87	1.0	0.0	NaN	0
166	1.0	NaN	3.0	0
192	2.0	NaN	7.0	1
266	2.0	0.0	NaN	1
287	2.0	NaN	7.0	0
302	1.0	NaN	3.0	0

Figure 13 Visualisation des lignes de valeurs manquantes.

Pour remédier à ce problème, nous avons choisi d'utiliser la technique de suppression des lignes contenant des valeurs manquantes (NaN).

5.2. Séparation de la Colonne Cible « num »

Pour intégrer le dataset dans un modèle d'apprentissage automatique (ML), il est crucial de séparer la colonne cible (variable dépendante) du reste des données (variables indépendantes). Ainsi, dans ce contexte, la colonne 'num' sera désignée comme la variable cible

Y, tandis que les autres colonnes seront utilisées comme variables caractéristiques X. Pour cela, nous avons séparé les données en deux ensembles : X (les caractéristiques) et Y (la cible). X contiendra toutes les colonnes sauf 'num', tandis que Y contiendra uniquement la colonne 'num'.

5.3. Normalisation des données

Cette étape a pour objectif d'uniformiser les entrées du modèle ‘X’ en les mettant à la même échelle, ce qui permet de les comparer et de les analyser plus efficacement. Pour ce faire, on utilise des techniques telles que la mise à l'échelle min-max ou la normalisation par l'écart-type (standardisation).

- Mise à l'échelle min-max : Cette technique ajuste les données pour qu'elles se situent dans une plage spécifique, généralement entre 0 et 1.
- Normalisation par l'écart-type (standardisation) : Cette méthode transforme les données pour qu'elles aient une moyenne de zéro (0) et un écart-type d'un (1), facilitant ainsi leur comparaison et leur analyse.

La normalisation garantit que toutes les variables ont un impact équitable sur le modèle. Nous avons choisi la standardisation pour notre modèle :

$$X_{\text{normalisé}} = \frac{X - \mu}{\sigma}$$

Où μ est la moyenne et σ est l'écart-type des données.

La Figure 14 (a) présente un échantillon des valeurs de ‘X’ après standardisation, tandis que la Figure 14 (b) montre la variable cible ‘Y’ dans son état brut.

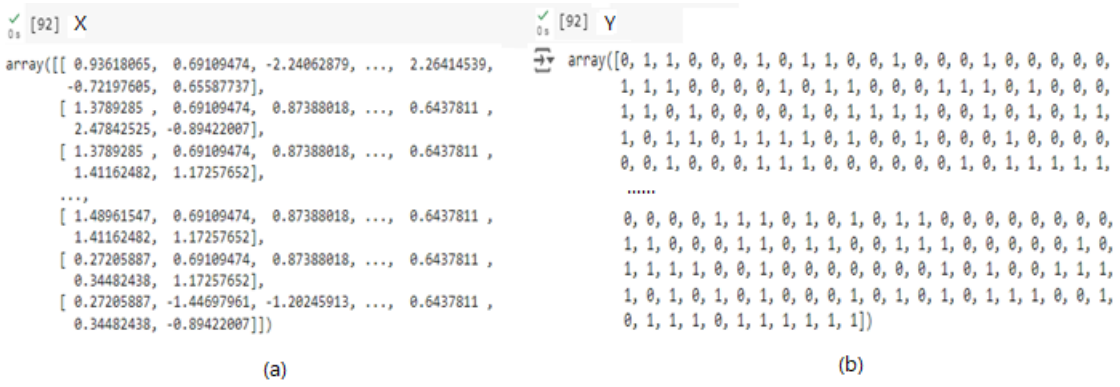

Figure 14 X après standardisation, (b) : Variable cible ‘Y’

6. Entrainement des modelés en appliquant les techniques d'apprentissage automatique

6.1. Création des ensembles de données d'entrainement et de test

Après la phase de prétraitement des données, nous avons divisé notre jeu de données en ensembles d'entraînement et de test. Concrètement, nous avons utilisé la fonction ***train_test_split( )*** de la bibliothèque Sklearn. Cette fonction nous a permis de séparer les données en deux groupes : 80 % des données ont été affectées à l'ensemble d'entraînement, destiné à ajuster notre modèle, et les 20 % restants ont été affectés à l'ensemble de test, utilisé pour évaluer la performance du modèle. Cette division permet d'assurer que notre modèle peut généraliser efficacement à de nouvelles données.

6.2. Apprentissage automatique

L'apprentissage automatique (ML) est un domaine de l'intelligence artificielle qui consiste à construire des algorithmes capables d'apprendre de l'expérience. La façon dont fonctionnent les algorithmes de ML est qu'ils détectent les motifs cachés dans l'ensemble de données en entrée et construisent des modèles. Ensuite, ils peuvent faire des prédictions précises pour de nouveaux ensembles de données qui sont entièrement nouveaux pour les algorithmes. [22]

Les catégories d'apprentissage automatique connues sous le nom d'apprentissage supervisé et d'apprentissage non supervisé sont essentielles dans les applications d'intelligence artificielle, jouant un rôle crucial dans la compréhension et le traitement des données. L'apprentissage supervisé se concentre sur le développement de modèles capables de prédire de nouveaux résultats à l'aide de données d'entraînement préalablement étiquetées. En revanche, l'apprentissage non supervisé analyse des données non étiquetées pour découvrir des structures ou des modèles cachés, comme dans le cas du clustering ou de la réduction de dimensionnalité.

L'une des applications les plus importantes dans ce contexte est la classification, où les modèles sont utilisés pour classifier les données en catégories spécifiques en fonction des caractéristiques préalablement connues. La classification est un outil essentiel dans un large éventail de domaines, notamment la classification d'images, la classification linguistique, la classification de documents, etc., contribuant à une meilleure compréhension des données et à

la prise de décisions basée sur les connaissances acquises. Parmi les algorithmes de classification les plus importants en apprentissage supervisé, on trouve :

- Les arbres de décision
- Les machines à vecteurs de support (SVM)
- Les forêts aléatoires
- Les réseaux de neurones
- Les k-plus proches voisins (k-NN)

Dans notre étude, nous allons utiliser la classification pour analyser et prédire la présence ou l'absence de maladies cardiaques à partir de notre jeu de données. En particulier, nous nous concentrerons sur la construction de modèles de classification qui nous permettront de catégoriser les données en fonction des caractéristiques identifiées lors de la phase de prétraitement. Cette approche nous aidera à tirer des conclusions précises et à améliorer la prise de décision clinique basée sur les données analysées.

6.2.1. Machines à vecteurs de support (SVM)

Les SVM, ou machines à vecteurs de support, sont des modèles d'apprentissage automatique utilisés pour la classification. Il vise t à trouver l'hyperplan optimal qui sépare les différentes classes dans un espace de dimension supérieure. En bref, les SVM sont un ensemble de techniques d'apprentissage supervisé visant à trouver, dans un espace de dimension supérieure à un, l'hyperplan optimal pour diviser un jeu de données en deux. Ces modèles fonctionnent comme des séparateurs linéaires, où la frontière entre les classes est une droite**.** [23]

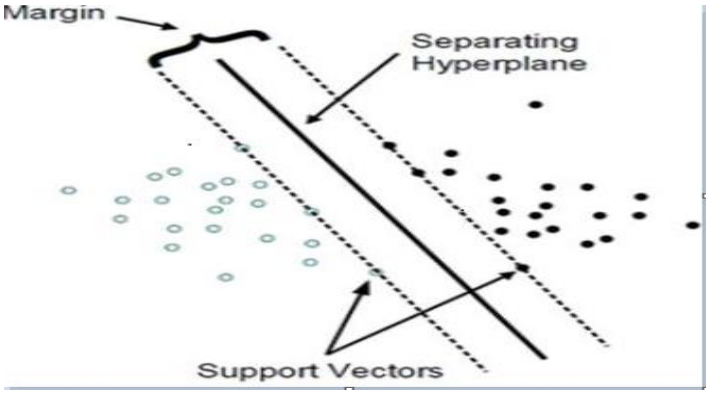

Figure 15 Machines à vecteurs de support. [24]

L'application du modèle SVM avec modification des hyper-paramètres a permis d'explorer différentes configurations et d'évaluer l'impact de ces variations sur les performances du modèle. En analysant les résultats obtenus, on observe une variation significative des métriques d'évaluation en fonction des valeurs des hyper-paramètres et du noyau utilisé.

Tout d'abord, lors de l'utilisation du noyau RBF avec C = 0.1 et Gamma = 0.1, on obtient une précision, un rappel et une F-mesure de 90%, ce qui indique une performance relativement équilibrée du modèle. En revanche, en utilisant le noyau sigmoid avec les mêmes valeurs de C et Gamma, on observe une amélioration notable des performances avec une précision, un rappel et une F-mesure de 93.44%, 93.33%, et 93.27% respectivement, suggérant que ce noyau est plus adapté à notre jeu de données.

En utilisant le noyau linéaire avec les mêmes valeurs des hyper-paramètres, les performances restent stables avec une précision, un rappel et une F-mesure de 90%. Cela suggère que pour ces données, un modèle linéaire fournit des performances comparables à celles obtenues avec des noyaux plus complexes.

Enfin, l'utilisation du noyau poly avec C = 0.1 et Gamma = 0.1 conduit à une baisse significative des performances, avec une précision de 78.33%, un rappel de 81.60%, et une F-mesure de 76.54%. Cela indique que ce noyau et ces valeurs d'hyper-paramètres ne sont pas bien adaptés aux caractéristiques des données.

Les résultats obtenus sont résumés dans le tableau 3.

Métriques d'évaluation / Hyper-Paramétres	Accuracy(%)	Precision (%)	Recall (%)	F-measure(%)
Kernel = rbf C = 0.1 Gamma = 0.1	90%	90%	90%	90%
Kernel = sigmoid **C = 0.1** **Gamma = 0.1**	**93.33%**	**93.44%**	**93.33%**	**93.27%**
Kernel = linear C = 0.1 Gamma = 0.1	90%	90%	90%	90%
Kernel = poly C = 0.1 Gamma = 0.1	78.33%	81.60%	78.33%	76.54%

Tableau 3 Evaluation des résultats selon l'hyper-paramètre « Kernel.«

- **La matrice de cofusion du modèle SVM**

La matrice de confusion du modèle SVM présentée dans la figure 16 offre un aperçu détaillé des performances de classification. Cette matrice permet de visualiser le nombre de prédictions correctes et incorrectes en comparant les valeurs réelles aux valeurs prédites.

La matrice de confusion démontre que le modèle SVM est très efficace pour la classification avec un taux de précision élevé et une bonne performance globale. Cependant, il y a encore des erreurs de classification, notamment quelques faux négatifs et un faux positif, indiquant qu'il y a un léger biais vers la classe majoritaire.

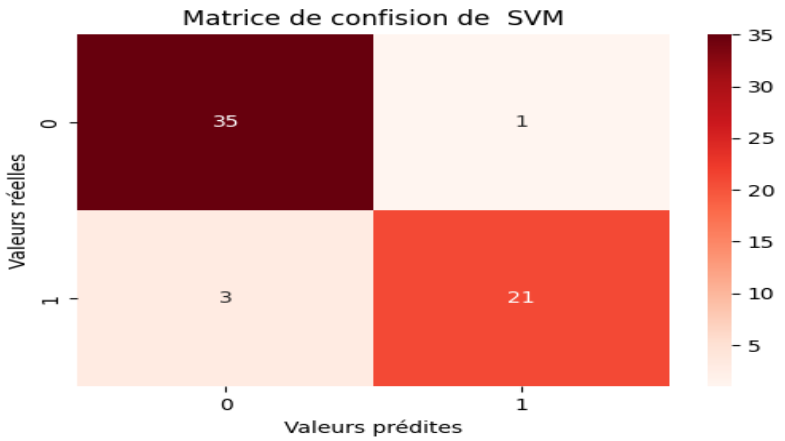

Figure 16 Matrice de confusion du modèle SVM.

6.2.2. Forêt aléatoire (RF)

Le modèle Forêt aléatoire, Random Forest (RF), est un algorithme d'apprentissage automatique populaire qui appartient à la technique d'apprentissage supervisé. Il peut être utilisé pour les problèmes de classification et de régression en ML. Il est basé sur le concept d'apprentissage d'ensemble, qui est un processus de combinaison de plusieurs classificateurs pour résoudre un problème complexe et améliorer les performances du modèle. [25]

- Fonctionne l'algorithme RF

Le processus de travail peut être expliqué dans les étapes et le diagramme ci-dessous :

- **Étape 1 :** Sélectionnez des points de données K aléatoires dans l'ensemble d'entraînement.
- **Étape 2 :** Créez les arbres de décision associés aux points de données sélectionnés (sous-ensembles).
- **Étape 3 :** Choisissez le numéro N pour les arbres de décision que vous souhaitez créer.
- **Étape 4 :** Répétez l'étape 1 et 2.
- **Étape 5 :** Pour les nouveaux points de données, recherchez les prédictions de chaque arbre de décision et attribuez les nouveaux points de données à la catégorie qui remporte les votes majoritaires.

Le diagramme ci-dessous explique le fonctionnement de l'algorithme RF.

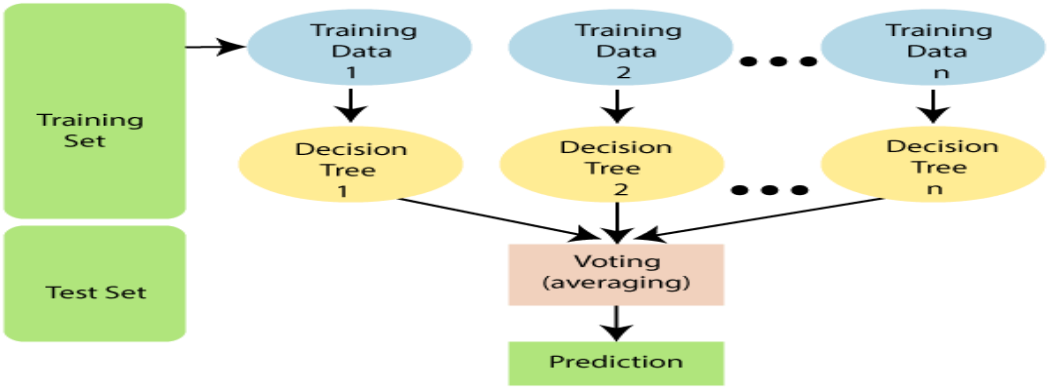

Figure 17 Fonctionnement de l'algorithme RF. [25]

L'application du modèle Random Forest (RF) sur notre jeu de données a permis d'évaluer l'impact du nombre d'estimateurs sur les performances du modèle. Les résultats sont résumés dans le tableau 4, mettant en évidence les variations des métriques d'évaluation en fonction des différents hyper-paramètres.

Métriques d'évaluation / Hyper-Paramétres	Accuracy(%)	Precision(%)	Recall(%)	F-measure(%)
Estimateurs=8	88.33%	89.19%	88.33%	88.03%
Estimateurs=10	**90%**	**90.04%**	**90%**	**89.91%**
Estimateurs=75	85%	84.93%	85%	84.94%
Estimateurs=100	88.33%	88.45%	88.33%	88.37%

Tableau 4 Evaluation des résultats par Estimateurs

- Avec 8 estimateurs, le modèle RF affiche une précision et un rappel relativement équilibrés, avec une F-mesure de 88.03%, indiquant une bonne performance globale.
- En augmentant le nombre d'estimateurs à 10, les performances du modèle s'améliorent légèrement, avec une précision et un rappel de 90%, et une F-mesure de 89.91%,

suggérant que ce paramétrage offre la meilleure performance parmi les configurations testées.

- Avec 75 estimateurs, on observe une diminution des performances globales, avec une précision et un rappel de 85% et une F-mesure de 84.94%. Cela peut indiquer un surajustement ou une saturation des bénéfices apportés par l'augmentation du nombre d'estimateurs.
- En utilisant 100 estimateurs, les performances se stabilisent, avec des valeurs de précision et de rappel similaires à celles obtenues avec 8 estimateurs, et une F-mesure de 88.37%.

- **La matrice de cofusion du modèle RF**

La matrice de confusion présentée dans la Figure 18 démontre que le modèle RF offre une bonne performance de classification avec une précision élevée de 90%. Toutefois, il y a quelques erreurs de classification, notamment quatre faux négatifs et deux faux positifs.

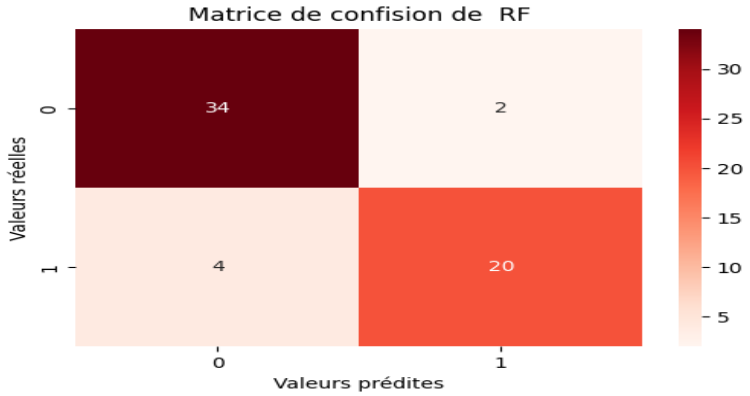

Figure 18 Matrice de confusion du modèle RF.

6.2.3. K plus proches voisins (KNN)

L'algorithme des K plus proches voisins ou K-nearest neighbors (KNN) est un algorithme de Machine Learning qui appartient à la classe des algorithmes d'apprentissage supervisé simple et facile à mettre en œuvre qui peut être utilisé pour résoudre les problèmes de classification et de régression [26].

L'intuition derrière l'algorithme des K plus proches voisins est l'une des plus simples de tous les algorithmes de Machine Learning supervisé 7 :

- **Étape 1 :** Sélectionnez le nombre K de voisins.
- **Étape 2 :** Calculez la distance.
- **Étape 3 :** Prenez les K voisins les plus proches selon la distance calculée.
- **Étape 4 :** Parmi ces K voisins, comptez le nombre de points appartenant à chaque catégorie.
- **Étape 5 :** Attribuez le nouveau point à la catégorie les plus présents parmi ces K voisins.
- **Étape 6 :** Notre modèle est prêt.

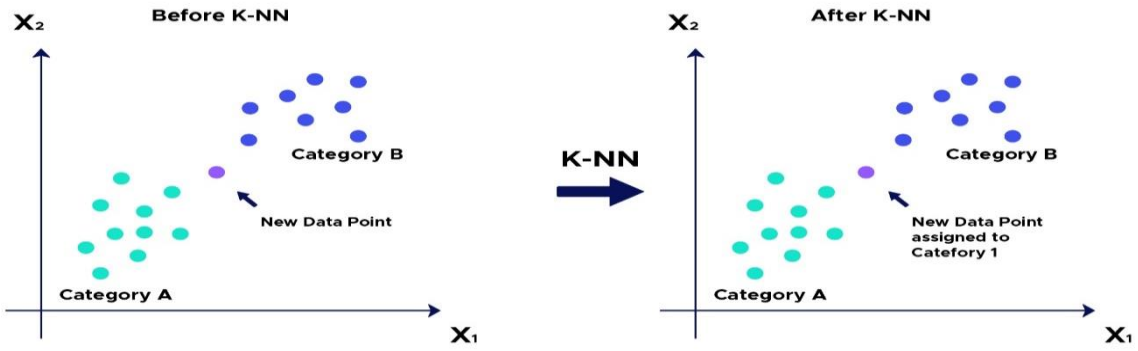

Figure 19 K plus proches voisins (KNN). [26]

L'application du modèle KNN sur notre jeu de données a permis d'évaluer l'impact de la sélection de différentes valeurs pour le paramètre N_neighbors. Les résultats obtenus sont illustrés dans le tableau 5.

Métriques d'évaluation / Hyper-Paramétres	Accuracy(%)	Precision(%)	Rcall(%)	F-measure(%)
N_neighbors=6	83.33%	83.49%	83.33%	83.01%
N_neighbors =7	85%	84.93%	85%	84.94%
N_neighbors =10	86.66%	86.65%	86.66%	86.55%
N_neighbors =20	**88.33%**	**88.45%**	**88.33%**	**88.37%**

Tableau 5 Evaluation des résultats par N_neighbors.

- Avec 6 voisins, le modèle KNN affiche une précision et un rappel équilibrés, avec une F-mesure de 83.01%, indiquant des performances correctes.
- En augmentant le nombre de voisins à 7, les performances du modèle s'améliorent légèrement avec une précision et un rappel de 85%, et une F-mesure de 84.94%.
- Avec 10 voisins, les performances continuent de s'améliorer, avec une précision et un rappel de 86.66%, et une F-mesure de 86.55%, suggérant que ce paramétrage est optimal pour le modèle KNN.

Avec 20 voisins, le modèle atteint sa meilleure performance, avec une précision de 88.33%, un rappel de 88.33%, et une F-mesure de 88.37%.

- **La matrice de cofusion du modèle KNN (K=20)**

La matrice de confusion présentée dans la Figure 20 démontre que le modèle KNN offre une bonne performance de classification avec une précision élevée de 88.33%. Cependant, il y a quelques erreurs de classification, notamment quatre faux positifs et trois faux négatifs.

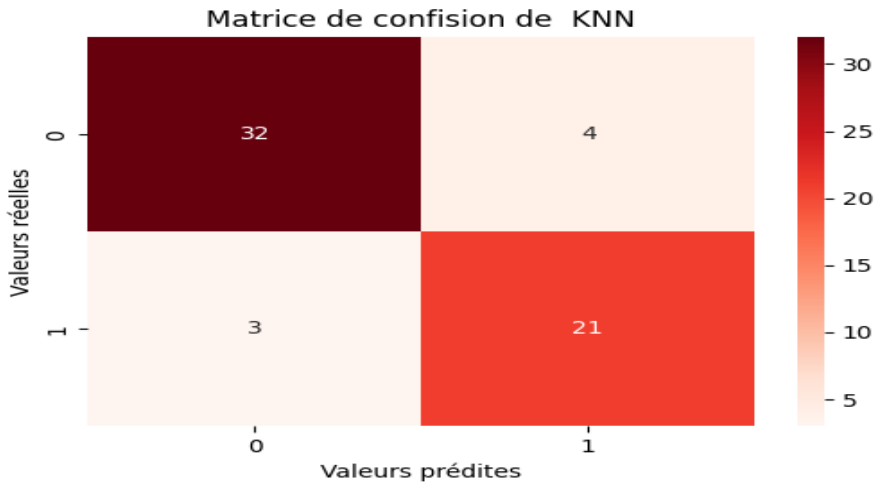

Figure 20 Matrice de confusion du modèle KNN.

6.2.4. Arbre de décision (DT)

Un arbre de décision est une structure arborescente de type organigramme dans laquelle chaque nœud interne désigne la fonctionnalité, les branches désignent les règles et les nœuds feuilles désignent le résultat de l'algorithme. Il s'agit d'un algorithme d'apprentissage automatique supervisé polyvalent, utilisé à la fois pour les problèmes de classification et de régression. C'est l'un des algorithmes les plus puissants. Et il est également utilisé dans Random Forest pour s'entraîner sur différents sous-ensembles de données d'entraînement, ce qui fait de Random Forest l'un des algorithmes les plus puissants de l'apprentissage automatique . [27]

- **Fonctionne l'algorithme de l'arbre de décision**

L'arbre de décision fonctionne en analysant l'ensemble de données pour prédire sa classification. Cela commence au nœud racine de l'arborescence, où l'algorithme visualise la valeur de l'attribut racine par rapport à l'attribut de l'enregistrement dans l'ensemble de données réel. Sur la base de la comparaison, il suit la branche et passe au nœud suivant. L'algorithme répète cette action pour chaque nœud suivant en comparant ses valeurs d'attribut avec celles des sous-nœuds et en poursuivant le processus. Il se répète jusqu'à ce qu'il atteigne le nœud feuille de l'arbre. Le mécanisme complet peut être mieux expliqué grâce à l'algorithme donné ci-dessous. [27]

- **Étape 1** : Commencez l'arborescence avec le nœud racine, dit S, qui contient l'ensemble de données complet.
- **Étape 2** : Trouvez le meilleur attribut de l'ensemble de données à l'aide de la mesure de sélection d'attribut (ASM).
- **Étape 3** : Divisez le S en sous-ensembles contenant les valeurs possibles pour les meilleurs attributs.
- **Étape 4 :** générez le nœud de l'arbre de décision, qui contient le meilleur attribut.
- **Étape 5 :** Créez de manière récursive de nouveaux arbres de décision en utilisant les sous-ensembles de l'ensemble de données créé à l'étape -3. Continuez ce processus jusqu'à ce qu'une étape soit atteinte où vous ne pouvez pas classer davantage les nœuds et appelez le nœud final en tant qu'algorithme d'arbre de classification et de régression de nœud feuille.

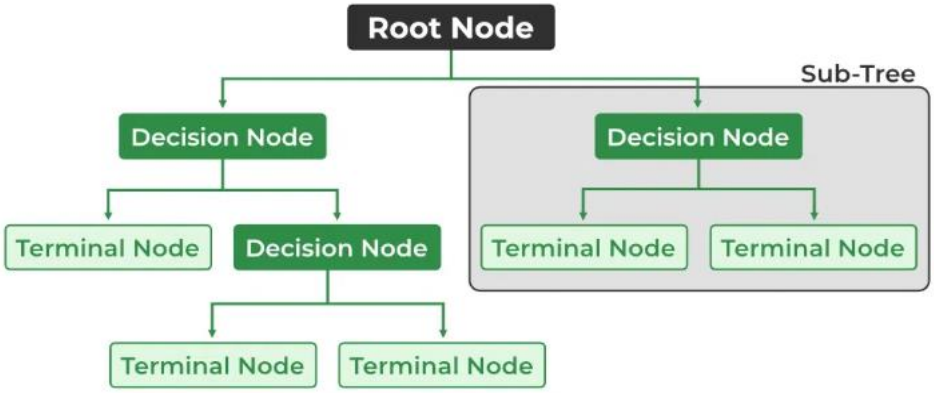

Figure 21 Arbre de décision (DT) [27]

L'application du modèle DT sur notre jeu de données a permis d'évaluer l'impact de la sélection de différentes valeurs pour l'hyper-paramètre Max_depth. Les résultats obtenus sont illustrés dans le tableau 6.

Métriques d'évaluation / Hyper-Paramétres	**Accuracy(%)**	**Precision(%)**	**Recall(%)**	**F-measure(%)**
Max_depth=5	83.33%	83.71%	83.33%	83.42%
Max_depth=7	80%	80.40%	80%	80.11%
Max_depth=8	83.33%	84.37%	83.33%	83.48%
Max_depth=10	**83.33%**	**84.37%**	**83.33%**	**83.48%**

Tableau 6 Evaluation des résultats par Max_depth.

- À une profondeur maximale de 5, le modèle affiche une précision et un rappel équilibrés avec une F-mesure de 83.42%, indiquant une bonne performance globale.

- En augmentant la profondeur maximale à 7, on observe une légère diminution des performances avec une précision et un rappel de 80% et une F-mesure de 80.11%. Cela peut indiquer que le modèle commence à surajuster les données.
- Avec une profondeur maximale de 8, les performances s'améliorent légèrement par rapport à une profondeur de 5, avec une précision de 84.37% et une F-mesure de 83.48%. Cela suggère que le modèle profite de l'augmentation de la profondeur pour mieux capturer les relations dans les données.
- À une profondeur maximale de 10, les performances restent constantes par rapport à une profondeur de 8, avec des valeurs identiques pour toutes les métriques. Cela indique que, au-delà d'une certaine profondeur, augmenter la complexité du modèle n'apporte pas de bénéfices supplémentaires et pourrait potentiellement entraîner une sur-adaptation.

- **La matrice de cofusion du modèle DT (DT=10)**

La matrice de confusion présentée dans la Figure 22 démontre que le modèle DT offre une bonne performance de classification avec une précision élevée de 84.37%. Cependant, il y a quelques erreurs de classification, notamment sept faux positifs et trois faux négatifs.

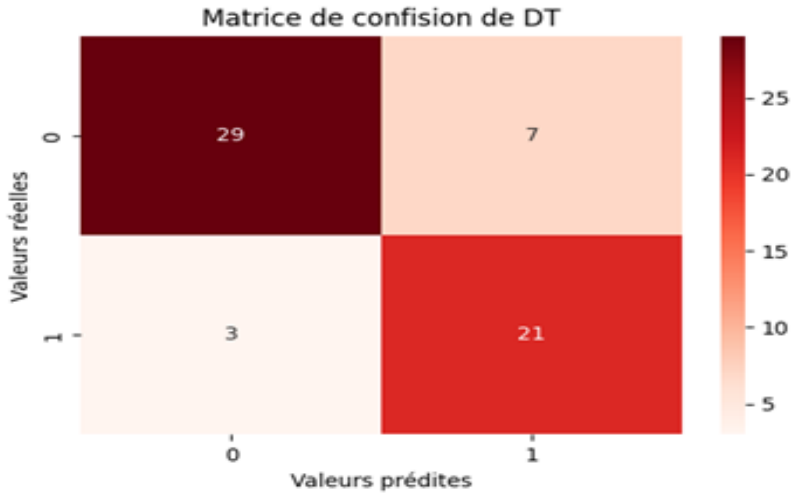

Figure 22 Matrice de confusion de l'algorithme DT.

6.2.5. Analyse des résultats obtenus suite à l'application des différents modèles

Le tableau 7 présente une évaluation comparative des performances de divers modèles de classification, incluant SVM, RF, KNN et DT. Les métriques de performance sont Accuracy, Precision, Recall et F-measure.

Métriques d'évaluation / Model	Accuracy(%)	Precision(%)	Recall(%)	F-measure(%)
SVM Kernel= sigmoid	**93.33%**	**93.44%**	**93.33%**	**93.27%**
RF Estimateur= 10	90%	90.04%	90%	89.91%
KNN N_neighbors=20	88.33%	88.45%	88.33%	88.37%
DT Max_depth9=10	83.33%	84.37%	83.33%	83.48%

Tableau 7 Evaluation des résultats des modelés utilisés.

- Le modèle SVM avec un noyau sigmoid obtient les meilleures performances globales parmi tous les modèles testés. Avec une précision de 93.33%, il montre une excellente capacité à classifier correctement les instances. Les valeurs élevées de précision, rappel et F-mesure indiquent que ce modèle est à la fois précis et capable de capturer efficacement les classes positives.
- Le modèle Random Forest, avec 10 estimateurs, présente également de très bonnes performances avec une précision de 90%. Il offre un bon équilibre entre précision et rappel, comme le montre sa F-mesure de 89.91%. Cela indique que ce modèle est robuste et bien équilibré.
- Le modèle KNN avec K=20 affiche une précision de 88.33%. Ses performances sont légèrement inférieures à celles du SVM et du RF, mais restent néanmoins solides. La F-mesure de 88.37% indique une bonne cohérence entre précision et rappel, ce qui montre que KNN est également un modèle fiable.

- Le modèle Decision Tree, avec une profondeur maximale de 10, obtient la précision la plus faible parmi les modèles testés, avec une valeur de 83.33%. Bien que ses métriques de précision et de rappel soient respectables, il est clair qu'il ne performe pas aussi bien que les autres modèles.

7. Conclusion

Dans ce chapitre, nous avons comparé plusieurs modèles de classification pour analyser notre jeu de données, en utilisant le Support Vector Machine (SVM), le Random Forest (RF), le K-Nearest Neighbors (KNN) et le Decision Tree (DT). Les résultats ont montré que le SVM avec un noyau sigmoid a offert les meilleures performances avec une précision de 93.33%, suivi de près par le Random Forest avec 10 estimateurs, qui a atteint une précision de 90%. Le KNN avec 20 voisins a également obtenu de bons résultats avec une précision de 88.33%. En revanche, le Decision Tree avec une profondeur maximale de 10 a présenté les performances les plus faibles avec une précision de 83.33%. Ces observations soulignent l'importance de choisir et d'optimiser correctement les modèles de classification.

Le SVM est recommandé comme modèle principal, mais le Random Forest reste une option solide pour des applications nécessitant une robustesse supplémentaire. Pour améliorer les performances, il est suggéré d'ajuster davantage les hyper-paramètres, d'enrichir le jeu de données et d'utiliser des techniques de régularisation. En conclusion, ce chapitre a mis en évidence les forces et faiblesses des différents modèles, fournissant des indications précieuses pour les futures applications cliniques.

Chapitre 3 :

Sélection de Caractéristiques par Application d'un Algorithme Génétique

1. Introduction

La sélection des caractéristiques est essentielle pour traiter les ensembles de données de grande dimension, souvent riches en informations mais également encombrés de caractéristiques redondantes ou non pertinentes. Ce processus vise à identifier les caractéristiques les plus pertinentes pour améliorer l'efficacité et la précision des modèles d'apprentissage automatique.

Dans notre contexte spécifique, la réduction du nombre d'attributs présente deux finalités principales : premièrement, elle vise à diminuer le coût des tests nécessaires pour la prédiction des maladies cardiaques. Deuxièmement,elle cherche à alléger le nombre d'attributs dans l'interface de l'application de prédiction une fois que le modèle de prédiction est achevé.

Parmi les diverses méthodes disponibles, les algorithmes génétiques se distinguent par leur capacité à explorer efficacement de vastes espaces de solutions. Inspirés par la sélection naturelle, ces algorithmes évoluent génération après génération pour identifier les sous-ensembles de caractéristiques optimaux.

Ce chapitre explore l'application des algorithmes génétiques pour la sélection des caractéristiques. Nous commencerons par une brève présentation des concepts de base de la réduction de dimensionnalité, suivie par une description des algorithmes génétiques et leur utilisation spécifique pour sélectionner les caractéristiques pertinentes dans notre cas d’étude.

2. Technique de réduction de dimensionnalité

Les ensembles de données comportant de nombreuses caractéristiques sont appelés données de grande dimension. Ils contiennent souvent beaucoup d'informations redondantes, y compris des facteurs liés ou dupliqués. La réduction de dimension vise à éliminer ces interférences. La réduction de la dimensionnalité des caractéristiques utilise les paramètres de caractéristiques existants pour former un espace de caractéristiques de faible dimension et surmonte les effets des informations redondantes ou non pertinentes, afin de mapper les informations efficaces contenues dans les caractéristiques d'origine vers un nombre réduit de caractéristiques. [28]

Dans un sens mathématique, supposons qu'il existe un vecteur de n dimension

$$X\ [x_1, x_2, x_3, .., x_n,]^T \quad (\mathbf{1})$$

X est transformé en un vecteur Y de m dimensions grâce à une fonction f, où

$$Y\ [y_1, y_2, y_3, .., y_m,]^T \quad (\mathbf{2})$$

Et $m << n$.

Le vecteur Y devrait contenir les principales caractéristiques du vecteur X. Mathématiquement, la fonction de mapping peut être exprimée comme suit : Ceci est le processus d'extraction et de sélection des caractéristiques. Il peut également être appelé le processus de "réduction de la perte de dimension" des données originales. Un vecteur de faible dimension résultant de la réduction de la dimension peut être appliqué aux domaines de la reconnaissance de motifs, de l'extraction de données et de l'apprentissage automatique. Cette fonction de mapping *f* est l'algorithme que nous voulons trouver pour la réduction des caractéristiques. Le choix du mapping *f* diffère en fonction du problème en attente. [28]

La réduction de dimensionnalité se divise en deux parties : la sélection des caractéristiques et l'extraction des caractéristiques et nous nous concentrerons sur cette partie sélection des caractéristiques. [28]

2.1. Sélection des caractéristiques

La sélection des caractéristiques, peut également être appelée sélection de variables ou sélection de sous-ensembles de caractéristiques, est un processus de sélection de sous-ensembles de caractéristiques qui sont appliqués à la construction du modèle [12]. Il y a quatre raisons pour l'utilisation des techniques de sélection des caractéristiques : simplifier le modèle pour le rendre plus facile à interpréter pour les chercheurs (utilisateurs) ; raccourcir le temps d'exécution ; éviter les maux de la dimensionnalité ; améliorer la généralisation en réduisant l'ajustement excessif (en réduisant formellement la variance). [28]

Le prérequis le plus important pour l'utilisation des techniques de sélection des caractéristiques est que les données contiennent de nombreuses caractéristiques redondantes ou liées qui peuvent être supprimées sans perdre beaucoup d'informations. Les caractéristiques redondantes ou liées sont deux concepts différents, car une caractéristique liée peut être redondante en présence d'autres caractéristiques liées qui lui sont étroitement associées. La

sélection des caractéristiques est généralement utilisée dans les domaines où il y a de nombreuses caractéristiques et relativement peu d'échantillons. [28]

Il est prouvé que c'est un problème NP. Seule la recherche exhaustive peut trouver la solution optimale. Le processus de recherche exhaustive a généralement un coût de calcul relativement élevé. Pour trouver le sous-ensemble optimal de caractéristiques, il faut rechercher les combinaisons de M caractéristiques parmi toutes les caractéristiques originales possibles N. Cette explosion de combinaisons conduit à une augmentation exponentielle du calcul avec l'augmentation du nombre total de caractéristiques. [28]

2.2. les avantages de de sélection des caractéristiques

La sélection des fonctionnalités, une phase essentielle du pipeline d'apprentissage automatique (ML), est le processus de sélection des variables ou des fonctionnalités les plus pertinentes dans un ensemble de données à utiliser pour la formation du modèle. Le processus offre de nombreux avantages en matière de développement de modèles et d'optimisation des performances. Il joue un rôle essentiel dans le prétraitement des données, transformant les ensembles de données brutes en entrées raffinées propices à un apprentissage de modèles précis et fiable. La sélection de fonctionnalités améliore les modèles ML en facilitant l'interprétabilité, en réduisant le surajustement, en améliorant la précision et en réduisant les coûts de calcul. [29]

- **Interprétabilité améliorée du modèle**

L'inclusion de fonctionnalités non pertinentes dans un modèle ML peut le rendre complexe, difficile à interpréter et peu fiable. La sélection des fonctionnalités simplifie les modèles en supprimant les fonctionnalités sans importance ou redondantes, les rendant ainsi plus compréhensibles. Cela peut être particulièrement bénéfique dans les domaines où l'interprétabilité est cruciale, comme la santé ou la finance, où les prédictions nécessitent souvent des explications aux parties prenantes ou aux organismes de réglementation. Un modèle plus facile à expliquer permet également aux data scientists d'obtenir de meilleures informations, leur permettant ainsi de l'affiner plus efficacement. [29]

- **Surapprentissage réduit**

Le surajustement est un problème courant en ML, où un modèle apprend trop bien les données d'entraînement, y compris le bruit et les valeurs aberrantes, ce qui entraîne une

mauvaise généralisation aux données invisibles. Lorsque des fonctionnalités non pertinentes sont présentes, un surajustement est plus susceptible de se produire en raison d'une complexité accrue du modèle. La sélection de fonctionnalités réduit le risque de surajustement en minimisant la complexité du modèle, améliorant ainsi sa capacité à généraliser à de nouvelles données. [29]

- **Précision améliorée**

La précision d'un modèle ML dépend largement de la qualité des données d'entrée. L'inclusion de fonctionnalités non pertinentes ou redondantes peut entraîner une diminution des performances du modèle en raison de données bruitées. La sélection des fonctionnalités atténue ce problème en garantissant que seules les fonctionnalités pertinentes, qui contribuent de manière significative au résultat, sont incluses. Cela conduit souvent à une augmentation de la précision du modèle, rendant les prévisions plus fiables et plus utiles. [29]

- **Coût de calcul réduit**

Les modèles ML peuvent devenir prohibitifs en termes de calcul et de ressources lorsqu'il s'agit de données de grande dimension. Chaque fonctionnalité supplémentaire peut augmenter considérablement le temps de formation et les besoins en mémoire. En identifiant et en conservant uniquement les caractéristiques significatives, la sélection des caractéristiques peut réduire considérablement la dimensionnalité de l'ensemble de données. Ceci, à son tour, accélère le processus de formation des modèles et réduit les besoins en mémoire, ce qui rend possible la formation de modèles sur des appareils dotés de ressources informatiques limitées. [29]

Dans cette section, nous présentons les algorithmes les plus importants associés à cette technique.

2.2.1. Algorithmes basés sur la stratégie de recherche

Les algorithmes basés sur la stratégie de recherche sont des techniques de sélection des caractéristiques qui explorent l'espace des possibles combinaisons de caractéristiques pour identifier les sous-ensembles les plus pertinents pour la construction de modèles prédictifs.

Ces stratégies peuvent être classées en deux catégories, la recherche exhaustive et la recherche métaheuristique.

2.2.1.1. Algorithme de recherche exhaustive

Cette méthode explore toutes les combinaisons possibles de caractéristiques pour identifier la meilleure. Bien que cette approche garantisse la découverte de la solution optimale, elle est souvent impraticable pour des ensembles de données volumineux en raison de son coût élevé.

- **Recherche par élagage et borne (BBS)**

L'algorithme de recherche par élagage et borne (Branch and Bound Search, BBS) est une méthode pour trouver des solutions optimales dans l'arbre de l'espace de solutions d'un problème. Il utilise généralement la méthode du coût minimal ou le parcours en largeur pour naviguer dans l'arbre. La principale idée derrière BBS est "l'élagage". Bien que BBS soit un algorithme d'optimisation exhaustif, l'ajout de bornes aux branches permet de réduire considérablement le nombre de scénarios à évaluer. [28]

Processus de BBS :

1. **Détermination des Bornes :** Les bornes supérieures et inférieures de la valeur cible sont établies au début.
2. **Expansion des Nœuds :** Lorsqu'un nœud est sélectionné pour expansion, tous ses nœuds enfants sont générés une seule fois.
3. **Élagage :** Les nœuds enfants qui ne peuvent pas contribuer à une solution optimale ou réalisable sont écartés. Seuls les nœuds prometteurs sont conservés dans la liste des nœuds actifs.
4. **Itération :** Le processus se répète en sélectionnant le prochain nœud dans la liste des nœuds actifs et en réitérant les étapes jusqu'à ce qu'une solution optimale soit trouvée ou que la liste des nœuds actifs soit vide. [28]

Cette approche permet de réduire le nombre de combinaisons à évaluer, rendant la recherche de solutions optimales plus pratique et moins coûteuse en termes de calcul.

2.2.1.2. Les métaheuristiques à base de population de solutions

Une grande variété de métaheuristiques basées sur une population de solutions a été proposée dans la littérature. Dans cette section, nous présentons deux exemples de ces algorithmes, à savoir les algorithmes génétiques et l'algorithme de colonies de fourmis.

- **Algorithme génétique (AG)**

L'algorithme génétique (AG) est une stratégie de recherche basée sur l'analogie de la théorie de la sélection naturelle. C'est un algorithme évolutif composé de quatre parties : un groupe d'individus (ou de chromosomes) qui peuvent représenter une solution possible ; une fonction appropriée pour évaluer la forme physique individuelle ; une fonction de sélection pour choisir l'individu qui est capable de produire la génération suivante ; il y a aussi un opérateur génétique, tel que le croisement et la mutation, pour explorer le nouvel espace de recherche. [28]

Chaque chromosome est associé à une mesure d'adaptation (fitness). En optimisation, cela correspond à la fonction objective. Un algorithme génétique sélectionne des paires de chromosomes parents, favorisant ceux avec une meilleure adaptation (fitness élevée), et génère de nouvelles solutions (enfants) en appliquant des opérateurs de croisement (crossover) et de mutation. L'objectif est que les bonnes solutions échangent leurs caractéristiques par croisement et produisent des solutions encore meilleures. Les principales étapes d'un algorithme génétique sont illustrées à la figure 23.

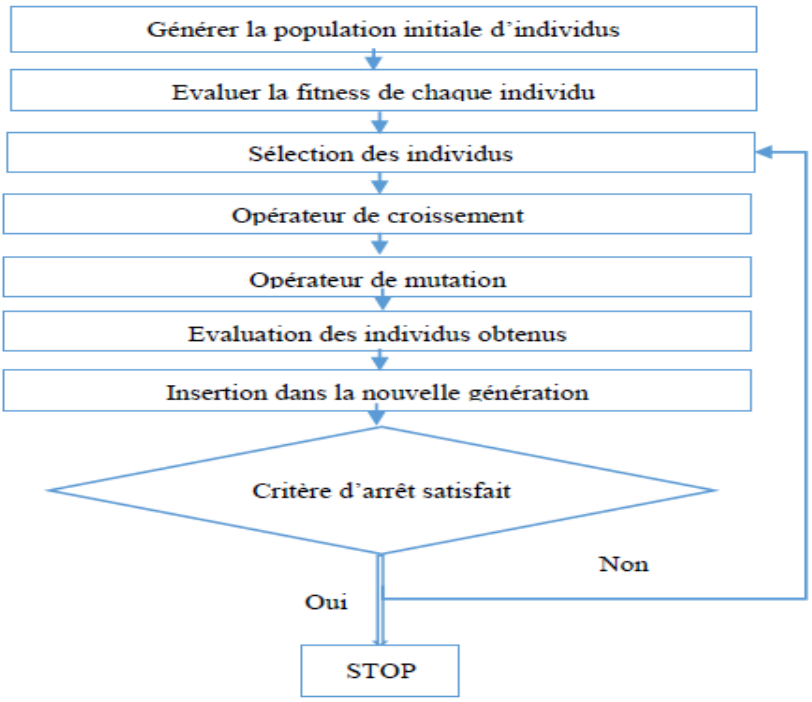

Figure 23 Diagramme de flux de l'algorithme génétique. [30]

- **Algorithme de colonies de fourmis (ACO)**

L'algorithme de colonies de fourmis (Ant Colony Optimization, ACO) est inspiré du comportement des fourmis dans leur recherche de nourriture. Les étapes de l'algorithme sont les suivantes :

1. La fourmi explore la région de manière aléatoire pour trouver de la nourriture.
2. Les fourmis ramènent la nourriture à la grotte et laissent des traces de phéromones chimiques.
3. La quantité de phéromones augmente avec l'augmentation de la quantité de nourriture.
4. Les autres fourmis trouvent les sources de nourriture en fonction de la trace de phéromones.

La première étape implique l'initialisation des traces de phéromones. Ensuite, selon les règles de transition d'état probabilistes, chaque fourmi crée une solution qui dépend de l'état des phéromones. Enfin, le nombre de phéromones change en deux phases : l'une est la phase d'évaporation, pendant laquelle une petite partie des phéromones s'évapore ; l'autre est la phase intensive, dans laquelle chaque fourmi possède un grand nombre de phéromones, et le nombre de phéromones est proportionnel à l'adaptabilité de la solution. Ce processus est itératif jusqu'à ce que les critères d'arrêt soient atteints. [28]

2.2.2. Basé sur le critère d'évaluation

Les techniques basées sur le critère d'évaluation jouent un rôle crucial dans le processus de sélection de caractéristiques, d'optimisation et de résolution de problèmes. Ces méthodes évaluent les solutions potentielles en utilisant des critères définis pour déterminer leur qualité ou performance.

- **Méthode de filtre**

Le filtre considère les critères utilise des ritères d'évaluation pour sélectionner les variables pertinentes. Le filtre trouve généralement une norme appropriée pour évaluer les variables et utilise un seuil pour éliminer les variables en dessous de ce seuil afin de filtrer les variables moins pertinentes, réduisant ainsi le degré de corrélation entre les caractéristiques et augmentant le degré de corrélation entre les caractéristiques et les classes. Le filtre est simple et pratique, et il est largement utilisé.

Les critères d'évaluation sont divisés en quatre catégories : basées sur la distance (distance euclidienne, distance de Mahalanobis, distance de Bhattacharyya, etc.), basées sur l'information (entropie de Shannon, entropie conditionnelle, gain d'information, information mutuelle, etc.), basées sur l'indépendance (pertinence, similitude) et basées sur la cohérence. [28]

- **Méthode de l'enveloppe**

L'enveloppe intègre le processus de sélection des caractéristiques dans l'apprentissage des algorithmes. Le prédicteur est considéré comme une boîte noire. Les performances de prédiction sont utilisées comme fonction objective pour évaluer le sous-ensemble de variables.

Contrairement à la méthode de filtre, la méthode de l'enveloppe est basée sur trois méthodes de composants : la stratégie de recherche, le prédicteur et la fonction d'évaluation. Le sous-ensemble de caractéristiques qui est évalué est déterminé par la stratégie de recherche. Le prédicteur peut être n'importe quelle méthode de classification. Ses performances sont utilisées comme fonction objective pour évaluer le sous-ensemble de caractéristiques déterminé par la stratégie de recherche afin de trouver le sous-ensemble optimal.

L'enveloppe est meilleure que le filtre, mais elle prend plus de temps et nécessite plus de ressources informatiques. [28]

3. Sélection des caractéristiques en applique algorithme génétique (AG)

La sélection des attributs constitue une étape fondamentale de notre étude. Cette section présente notre solution de sélection de caractéristiques pour le modèle de classification SVM, définie dans le chapitre précédent, en utilisant un algorithme génétique.

3.1.Description de la Solution

Le processus de sélection se déroule selon les étapes suivantes :

3.1.1. Préparation des données

Dans notre cas, nous avons utilisé l'ensemble de données préparé selon la section 5 du chapitre 2.

3.1.2. Fonction de Fitness

La fonction `getFitness` évalue la performance d'un individu (représentant un sous-ensemble de caractéristiques) en entraînant le modèle SVM et en calculant la Accuracy sur un ensemble de test.

3.1.3. Création des Individus et de la Population

Les opérations génétiques (mutation, croisement, sélection) ainsi que la création des individus sont définies. Deux paramètres doivent être initialisés pour l'algorithme génétique : la taille de la population et le nombre de générations. Dans notre cas, nous avons défini une population de 200 individus et exécuté l'algorithme sur 10 générations.

3.1.4. Sélection des Meilleurs Sous-ensembles

Les sous-ensembles de caractéristiques offrant la meilleure Accuracy de validation sont identifiés et affichés.

3.2.Analyse des résultats obtenus par l'algorithme génétique

L'algorithme génétique a permis de générer deux solutions :

- **Solution 01 :**

Accuracy avec toutes les caractéristiques	93.33%
Percentile	80.48%
Accuracy des caractéristiques sélectionnées	95%
Individus	[1,0,1,1,0,1,0,1,1,1,0,1,0]
Nombre de caractéristiques	8
Les caractéristiques sélectionnées	[‘age’, ‘cp’, ‘trestbps’, ‘fbs’, ‘thalach’, ‘exang’, ‘oldpeak’, ‘ca’]

Tableau 8 : Résultats de la Solution 01 (Accuracy et Sélection des Caractéristiques).

Dans ce cas, la sélection de 8 caractéristiques parmi toutes disponibles a permis d'améliorer l’exactitude de 93.33% à 95%. Le percentile de 80.48% indique que ce sous-ensemble est parmi les plus performants, mais pas le meilleur absolu. Ces résultats sont illustrés graphiquement dans la figure 24.

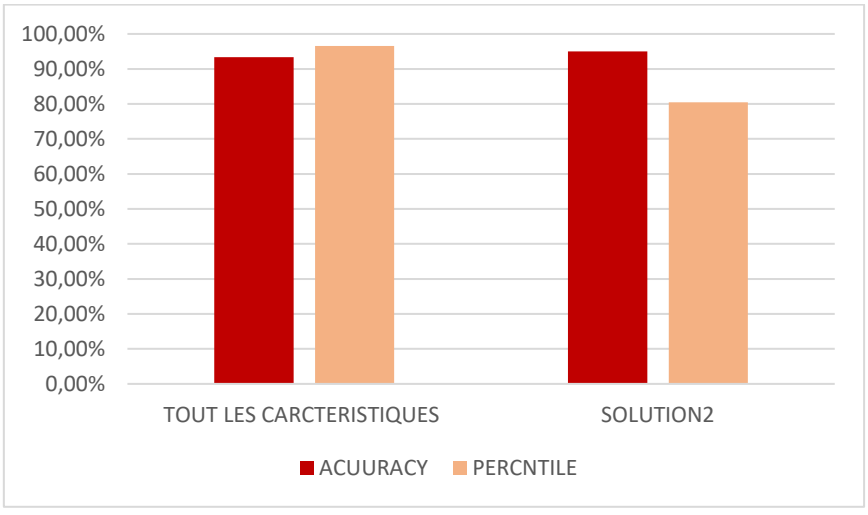

Figure 24 Solution N° 1 : Accuracy et percentile avant et après la sélection de caractéristiques.

- **Solution 02 :**

Accuracy avec toutes les caractéristiques	93.33%
Percentile	96.58%
Accuracy des caractéristiques sélectionnées	95%
Individus	[0,0,1,1,1,1,0,0,0,0,1,0,0]
Nombre de caractéristiques	5
Les caractéristiques sélectionnées	['cp', 'trestbps', 'chol', 'fbs', 'slope']

Tableau 9 : Résultats de la Solution 01 (Accuracy et Sélection des Caractéristiques)

Dans ce second cas, la sélection de seulement 5 caractéristiques a également permis d'améliorer l'accuracy à 95%. Le percentile de 96.58% montre que ce sous-ensemble est parmi les plus performants, presque au sommet en termes de précision. Ce résultat démontre que même avec un nombre réduit de caractéristiques, l'algorithme génétique peut trouver des combinaisons optimales qui surpassent l'utilisation de toutes les caractéristiques disponibles. Ces résultats sont illustrés graphiquement dans la figure 25.

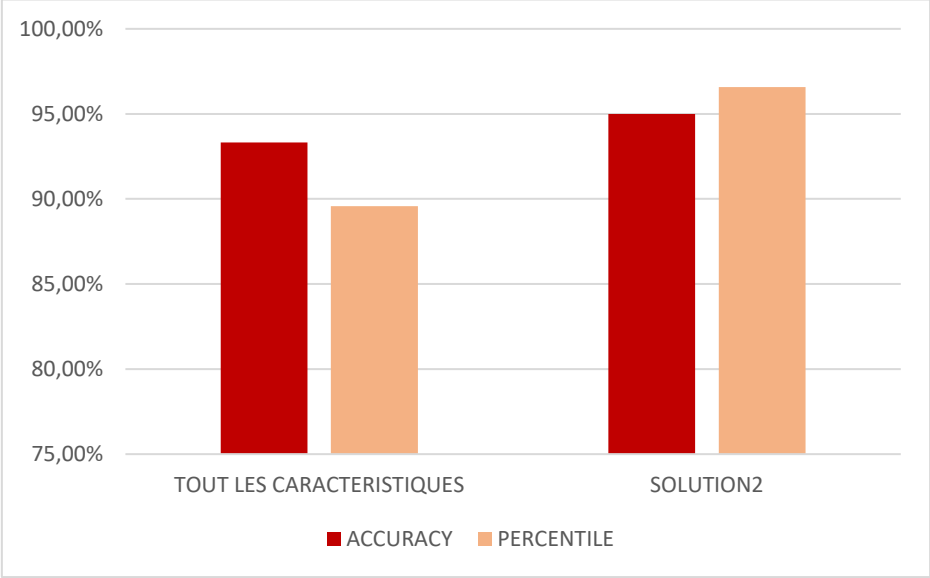

Figure 25 Solution N° 2 : Accuracy et percentile avant et après la sélection de caractéristiques.

La différence entre les deux solutions générées par l'algorithme génétique est illustrée par la représentation graphique de la Figure 26.

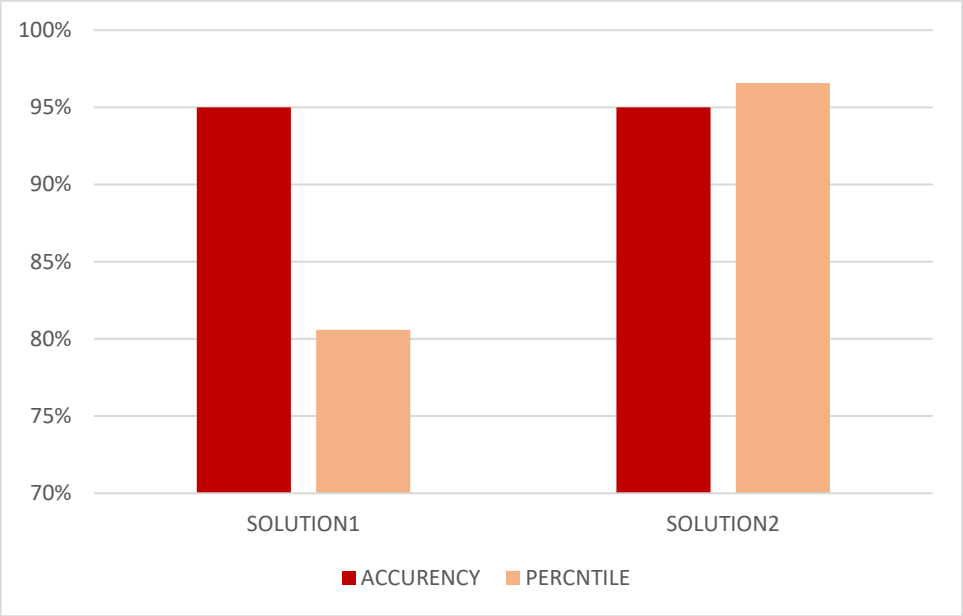

Figure 26 Représentation graphique de la comparaison entre la solution 1 et la solution 2

3.3.Discussion des résultats obtenus

En tenant compte des caractéristiques sélectionnées pour chaque cas, on peut dire que le deuxième cas est plus pertinent et réussi pour prédire les maladies cardiaques par rapport au premier cas. Les caractéristiques sélectionnées dans la deuxième solution, telles que le type de douleur thoracique, la tension artérielle au repos, le taux de cholestérol sérique, le taux de sucre dans le sang après un jeûne, et la pente du segment ST à l'exercice [‘cp’, ‘trestbps’, ‘chol’, ‘fbs’, ‘slope’], offrent une évaluation précise des symptômes et des changements physiologiques du cœur. Cette concentration sur les symptômes de la maladie et les facteurs physiologiques spécifiques aide grandement à déterminer la forme la plus précise pour estimer le risque de maladies cardiaques.

En revanche, la première solution se concentre sur des aspects plus généraux de la santé et de la performance cardiaque, ce qui peut le rendre moins apte à prédire avec précision les conditions cardiaques spécifiques par rapport à la deuxième solution. Ainsi, la deuxième solution est considérée comme plus efficace pour fournir une évaluation précise et spécifique des risques de maladies cardiaques.

De plus, les caractéristiques "cp", "trestbps", "chol", "fbs" et "slope" sont favorables à notre étude pour plusieurs raisons :

- **Facilité d'obtention :** Ces caractéristiques peuvent être mesurées facilement à l'aide de tests et d'examens standard effectués lors des visites médicales de routine.
- **Facilité de compréhension pour le patient :** Ces caractéristiques peuvent être expliquées facilement aux patients sans nécessiter de terminologie médicale complexe, ce qui facilite la compréhension et l'interprétation des résultats des tests pour les patients.
- **Non-onéreux pour le patient :** Ces caractéristiques ne nécessitent généralement pas de tests coûteux ou invasifs, ce qui les rend accessibles et abordables pour la plupart des patients.

4. Résumé des résultats des algorithmes d'apprentissage supervisé et de l'algorithme génétique

Dans cette section, nous résumerons toutes les résultats obtenus à partir de l'application des algorithmes d'apprentissage supervisé et de l'algorithme génétique. Après avoir analysé les données et évalué les performances, nous avons observé une amélioration significative dans l'exactitude des prédictions et l'efficacité de l'extraction des caractéristiques les plus influentes. Les résultats ont varié entre les différents algorithmes, certains montrant une meilleure capacité à manipuler les données et à améliorer les performances du modèle.

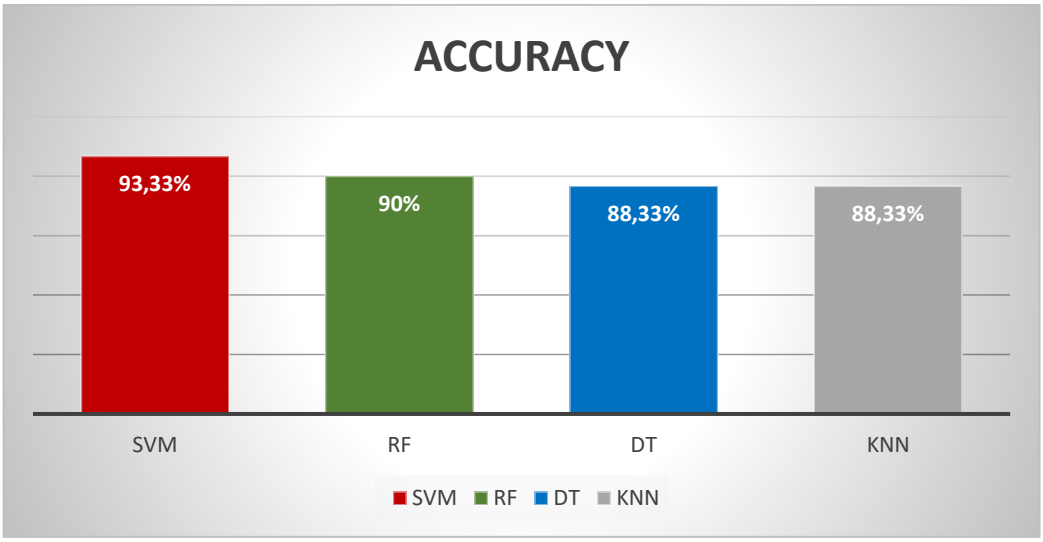

Figure 27 Résultats des algorithmes d'apprentissage supervisé.

Le Support Vector Machine (SVM) a été choisi comme algorithme de classification principal, suivi de l'application de l'algorithme génétique sur celui-ci. Cette étape vise à améliorer les performances du SVM en extrayant les caractéristiques les plus influentes et importantes pour la classification. Les résultats montrent que cette approche est en accord avec les données disponibles, ce qui a renforcé la capacité du SVM à généraliser et à améliorer son Accuracy.

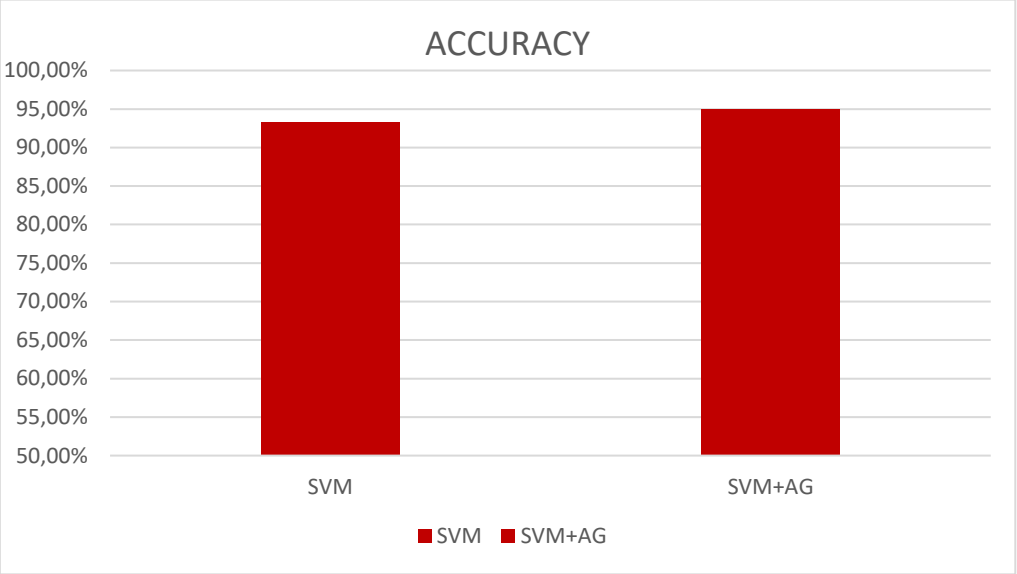

Figure 28 Résultats de comparaison entre SVM et SVM+AG.

L'application de l'algorithme génétique à SVM a conduit à une augmentation significative de l'Accuracy du modèle. Cette amélioration indique l'efficacité de l'algorithme génétique dans l'identification et la sélection des caractéristiques les plus pertinentes pour améliorer les performances du modèle SVM. Les résultats suggèrent que l'approche combinée de SVM et de l'algorithme génétique a permis de mieux adapter le modèle aux données disponibles, augmentant ainsi sa capacité à généraliser et à produire des prédictions plus précises.

5. Comparaison de notre modèle avec l'état de l'art

Pour avoir une bonne appréciation de notre modèle proposé, nous l'avons comparée aux différentes méthodes présentées dans notre revue de littérature. Les résultats obtenus sont présentés dans le tableau 10.

Référence	Dataset	Algorithmes	Métrique d'évaluation
Shadman Nashif et al ,2018	UCI Cleveland	**SVM** MLP Simple logistic NB RF	**SVM** Accuracy = 97.53% sensibilité = 97.5% Specificity = 94.94%
LIQAT ALI et al, 2019	UCI Cleveland	**SVM**	**SVM** Accuracy =92.22%
Gupta et al, 2022	UCI Cleveland	**LR** SVM NB DT KNN RF	**LR** Accuracy= 92.3%
Chiradeep Gupta et al, 2022	UCI Cleveland	**LR** SVM NB DT KNN RF	**LR** Sensibilité = 96.08% Spécificité = 87.5% Précision = 90.74% F1 Score = 93.34%
Notre approche (SVM + AG)	UCI Cleveland	**SVM** RF DT K-NN	**SVM** **Accuracy = 95%**

Tableau 10 : Performance des Algorithmes de Classification sur le Dataset UCI Cleveland

Les résultats obtenus montrent des performances de différents algorithmes de classification appliqués au dataset UCI Cleveland à travers diverses études. Shadman Nashif et al. (2018) ont obtenu des résultats exceptionnels avec SVM, atteignant une Accuracy de

97.53%, une sensibilité de 97.5% et une spécificité de 94.94%. En revanche, LIQAT ALI et al. (2019) ont rapporté une précision de 92.22% pour le SVM, inférieure à celle de Nashif et al, suggérant des variations potentielles dans les paramètres ou le prétraitement des données. Gupta et al. (2022) ont trouvé que la régression logistique (LR) était l'algorithme le plus performant avec une précision de 92.3%, tandis que Chiradeep Gupta et al. (2022) ont également mis en évidence la LR avec une sensibilité de 96.08%, une spécificité de 87.5%, une précision de 90.74% et un F1 Score de 93.34%. Enfin, notre approche (SVM+AG) a montré une précision de 95%, indiquant une amélioration significative par rapport à certaines études antérieures, bien que toujours légèrement en dessous des résultats de Shadman Nashif et al. Ces résultats montrent que bien que le SVM soit généralement performant, les résultats peuvent varier en fonction des configurations spécifiques et des techniques de prétraitement utilisées.

6. Conclusion

Dans la conclusion, nous soulignons l'importance de plusieurs points clés. Tout d'abord, nous avons abordé l'importance de l'extraction de caractéristiques, en mettant en évidence les défis et les avantages de cette étape cruciale dans le processus d'apprentissage automatique. Ensuite, nous avons examiné la sélection de caractéristiques et les algorithmes qui y sont associés, en mettant en évidence leurs diverses méthodes et leurs avantages respectifs dans la construction de modèles efficaces.

Un accent particulier a été mis sur l'algorithme génétique, où nous avons exploré son fonctionnement et son application dans le processus de sélection de caractéristiques. Nous avons montré soulignés comment cet algorithme peut conduire à des résultats optimaux en identifiant les caractéristiques les plus pertinentes pour un problème donné. Enfin, nous avons mis en lumière les résultats positifs obtenus grâce à l'application de l'algorithme génétique, montrant son efficacité dans l'amélioration des performances des modèles d'apprentissage automatique.

Conclusion générale

Les maladies cardiaques sont des conditions qui peuvent être prévenues ou dont les effets peuvent être réduits grâce à un diagnostic précoce. Étant donné la difficulté de prédire manuellement la probabilité de contracter des maladies cardiaques à partir de ce diagnostic, l'apprentissage automatique offre un outil efficace pour prendre des décisions et faire des prédictions en utilisant de grandes quantités de données contenant des informations détaillées sur les patients cardiaques et les personnes en bonne santé. Cette étude vise à développer un système intelligent pour améliorer la qualité des soins de santé et la prédiction des maladies cardiaques.

Pour atteindre notre objectif, nous avons utilisé l'ensemble de données Cleveland Heart Dataset, une base de données médicale contenant des facteurs de risque associés aux maladies cardiaques. Après une analyse et un traitement minutieux des données afin de les préparer à la classification, nous avons appliqué plusieurs algorithmes d'apprentissage supervisé, notamment KNN, RF, DT et SVM. Notre choix s'est porté sur le modèle SVM en raison de sa meilleure performance.

Ensuite, nous avons procédé à une étape de sélection de caractéristiques visant à améliorer la qualité de la prédiction des maladies cardiaques en réduisant le nombre de caractéristiques. Pour ce faire, nous avons utilisé l'algorithme génétique (AG), qui a permis de réduire la taille des données et du bruit, atteignant ainsi une précision de 95%.

Ce projet nous a été bénéfique à plusieurs égards. Sur le plan technique, il nous a permis de nous familiariser avec les concepts et les outils des sciences des données, du machine learning et de l'optimisation. Du point de vue pratique, nous avons pu exploiter l'environnement de développement Google Colaboratory, apprendre le langage de programmation Python ainsi que ses bibliothèques, et découvrir divers types de bases de données liées aux maladies cardiaques. Sur le plan culturel, nous avons acquis de nombreuses connaissances nouvelles spécifiques aux maladies cardiaques.

Quant aux perspectives, nous visons à :

- Intégrer l'apprentissage profond avec des algorithmes de sélection de caractéristiques afin d'améliorer la performance des modèles.
- Utiliser ou combiner d'autres ensembles de données en rapport avec les maladies cardiaque

Références

[1].**M, Pierre. 14 octobre 2015.** Généralités sur le cœur. https://www.sante-sur-le-net.com/maladies/cardiologie/generalites-coeur/. : s.n., 14 octobre 2015.

[2]. **Leem.** (n.d.). Maladies Cardiovasculaires (MCV).

[3]. Maladies cardiovasculaires. (Février 2024). Retrieved from https://www.who.int/fr/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[4]. Mayo, C. (Mars 2024). Maladie cardiaque. Retrieved from https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118

[5]. Facteurs de risque cardiovasculaire et prévention. (Avril 2024). Tests de maladies cardiaques.

[6]. Fondation des maladies du cœur et de l'AVC. (Mars 2024). Retrieved from Tests de maladies cardiaques: https://www.coeuretavc.ca/maladies-du-coeur/tests

[7]. PTB Diagnostic ECG Database. (Avril 2024). Retrieved from https://physionet.org/content/ptbdb/1.0.0/

[8]. Cardiac Atlas Project. (Avril 2024). Retrieved from https://www.cardiacatlas.org/mesa/

[9]. BHARDWAJ, A. (Avril 2024). Framingham heart study dataset. Retrieved from https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset

[10]. AVIGAN, A. (Mars 2023). Cleveland Clinic Heart Disease Dataset. Retrieved from https://www.kaggle.com/datasets/aavigan/cleveland-clinic-heart-disease-dataset

[11]. Ghosh, S. (Mars 2024). A Comprehensive Guide to Data Preprocessing. Retrieved from https://neptune.ai/blog/data-preprocessing-guide

[12]. Emilion, M. (03 mai 2024). Matrice de confusion : comment la lire et l'interpréter ? Retrieved from https://www.jedha.co/formation-ia/matrice-confusion

[13]. Matrice de confusion, la comprendre et l'utiliser. (Avril 2024). Retrieved from https://kobia.fr/classification-metrics-matrice-de-confusion/

[14]. Shadman Nashif, M. R. (n.d.). Heart Disease Detection by Using Machine. p.

[15]. LIAQAT ALI, A. N. (Mai 2019). An Optimized Stacked Support Vector Machines. https://ieeexplore.ieee.org/document/8684835

[16]. 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.

[17]. Gupta, C. (2022). Cardiac Disease Prediction using Supervised. pp. https://iopscience.iop.org/article/10.1088/1742-6596/2161/1/012013/pdf.

[18]. Angel Nancy, K. S. (July 2022). IoT-Cloud-Based Smart Healthcare Monitoring System for Heart Disease. pp. https://www.mdpi.com/2079-9292/11/15/2292.

[19]. A Angel Nancy, D. R.-Y. (15 June 2023). Fog-Based Smart Cardiovascular Disease Prediction System. pp. https://www.mdpi.com/2075-4418/13/12/2071.

[20]. (May 2024). Maladie cardiaque Cleveland

https://www.kaggle.com/datasets/ritwikb3/heart-disease-cleveland

[21]. Portail des connaissances statistiques. (May 2024). Retrieved from https://www.jmp.com/fr_fr/statistics-knowledge-portal/what-is-correlation.html

[22]. Maryam Aljanabi, M. H. (October 2018). Machine Learning Classification Techniques for Heart Disease Prediction: A Review. Maryam Aljanabi , Mahmoud H. Qutqut , Mohammad Hijjawi ; October 2018 Ihttps://www.researchgate.net/publication/328031918_Machine_Learning_Classification_Techniques_for_Heart_Disease_Prediction_A_Review.

[23]. Chrimni, W. (5 Septembre 2020). La revue IA. Retrieved from https://larevueia.fr/support-vector-machines-svm/

[24]. Mahesh, B. (January 2019). Machine Learning Algorithms -A Review. Retrieved from https://www.researchgate.net/publication/344717762_Machine_Learning_Algorithms_-A_Review

[25]. Random Forest Algorithm. (MAI 2024). Retrieved from https://www.javatpoint.com/machine-learning-random-forest-algorithm

[26] Qu'est ce que l'algorithme KNN ? (MAI 2024). https://datascientest.com/knn

[27]. Jain, A. (17 May, 2024). Decision Tree. https://www.geeksforgeeks.org/decision-tree/.

[28]. Weikuan Jia, M. S. (21 January 2022). Feature dimensionality reduction: a review. pp.

[29]. McCardel, B. (29May 2024). Feature Selection. pp. https://hex.tech/use-cases/feature-selection/.

[30]. Théses (PhD) ; Technologie Technologie > Génie Industriel ; École supérieure de technologie > Département de génie industriel ; 22 juillet 2016 à 10h17 ; http://eprints.univ-batna2.dz/id/eprint/93