


République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université AMO de Bouira

Faculté des Sciences et des Sciences Appliquées

Département d'Informatique

Mémoire de Master

en Informatique

Spécialité : *Ingénierie des systèmes d'informations et des logiciels*

Thème

Une approche hybride pour la prédiction des maladies cardiaques en utilisant le AG-KNN

Encadré par

— Mme BRAHIMI Farida

Réalisé par

— Mlle BEHLOULI Lina Aida

— Mr HAMDI Hichem

2021/2022

Remerciements

Tous d'abord, nous tenons à remercier Dieu clément et miséricordieux de nous avoir donné la force, la patience et le courage de mener à terme ce modeste travail. Comme, c'est un plaisir de remercier tous ceux qui ont contribué à la réalisation de ce mémoire :

En premier lieu, nous voudrons exprimer notre gratitude et nos remerciements les plus sincères à notre encadrante Mme BRAHIMI Farida, pour sa disponibilité, sa patience, son orientation, le soutien scientifique et moral qu'elle nous a apportés. vraiment avec elle nous avons eu le plaisir de travailler dans de bonnes conditions.

Nous tenons également à remercier Mme CHOUIREF Zahira pour les précieux conseils qu'elle nous a apportés durant la période de notre projet.

Nos vifs remerciements iront aussi aux membres de jury qui nous ferons l'honneur de juger notre travail et de soulever leurs critiques nécessaires permettront d'enrichir nos connaissances et d'apporter un plus à notre travail.

Nos remerciements sont destinés de même à tous nos enseignants du département informatique de l'université de Bouira Akli Mohand Oulhadj.

Dédicaces

Je dédie ce travail à :

À mes chers parents que nulle dédicace ne peut exprimer mes sincères sentiments, je les remercie pour leur patience illimitée, leurs encouragement continus ainsi que leur aide précieuse, en témoignage de mon profond amour et respect pour leur grands sacrifices.

À ma chère mère Rachida, pour tout ce qu'elle est pour moi, Que ce travail soit l'expression de ma reconnaissance pour votre soutient moral et matériel que vous n'avez cesse de prodiguer.

Vous avez tout fait pour mon bonheur et ma réussite. Que dieu vous préserve en bonne santé et vous accorde une longue vie.

À mes grands-mères que Dieux l'accueille dans son vaste paradis .

À mes chères Sœurs Anfel et Amina pour leurs soutiens et leurs encouragements .À mes chères amies Amira, Maroua et Asma pour leurs présence à mes cotés dans les moments les plus difficiles lors de la réalisation de ce travail, sans leurs encouragements ce travail n'aurait jamais vu le jour.

Et enfin à tous ceux qui m'ont donné la force de continuer et à tous ceux qui m'ont soutenue meme qu'avec un seul mot.

BEHLOULI Lina Aida.

Dédicaces

Avec tout respect et amour je dédie ce travail à :

A mes parents bien aimés, pour leur soutien inconditionnel et dont le seul objectif dans la vie est le bonheur et la satisfaction de leurs enfants. Je ne les remercierai jamais assez.

A mes frère : ali ,amine et sœur : karima ,naima et yasmine pour leurs encouragements et leur présence.

A tout mes amis notamment : lyza ,wissem ,kassem ,sidlai ,ayoub en souvenir des plus beaux instants qu'on passé ensemble .

Et en fin à tous ceux qui m'ont donné la force de continuer et à tous ceux qui m'ont soutenu, même si d'un seul mot.

HAMDI hichem.

Résumé

Les maladies cardiaques sont l'une des principales causes de décès dans le monde. Cependant, il reste difficile pour les cliniciens de prédire les maladies cardiaques car il s'agit d'une tâche complexe et coûteuse. C'est pourquoi, ces dernières années de nombreux chercheurs vont vers l'intelligence artificielle avec ses différentes méthodes d'apprentissage automatique et d'apprentissage profond pour prédire les maladies des personnes avant qu'elles ne surviennent.

Le but de cette étude est de proposer un système de prédiction des maladies cardiaques en utilisant l'algorithme d'apprentissage supervisé k-nearest neighbors (KNN) et gérer sa vulnérabilité par rapport au bruit des attributs non pertinents et redondants en utilisant l'algorithme génétique (AG). On l'a comparé avec d'autres approches d'apprentissage supervisées tel que Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), Suppor Vector Machine (SVM) dans le but de montrer sa performance. En utilisant les données médicales « Cleveland Heart Disease dataset CHDD » pour l'entrainement et le test du modèle.

Mots clés : *Les maladies cardiaques, Intelligence artificielle, Apprentissage automatique, Prédiction des maladies cardiaques, K plus proches voisins, Algorithme génétique, Ensemble de données sur les maladies cardiaques de Cleveland...*

Abstract

Heart disease is one of the leading causes of death in the world. However, it remains difficult for clinicians to predict heart disease because it is a complex and costly task. That is why, It is in recent years many researchers are moving towards artificial intelligence with its different methods of machine learning and deep learning to predict people's diseases before they occur. The purpose of this study is to propose a cardiac disease prediction system using the supervised learning algorithm k-nearest neighbors (KNN) and manage its susceptibility to noise from irrelevant and redundant attributes using the genetic algorithm (AG). It compared with other supervised learning approaches such as Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), Suppor Vector Machine (SVM) in order to show its performance. We used the Medical data "Cleveland Heart Disease dataset CHDD" for model training and testing.

Key words : *Heart Disease, Artificial Intelligence, Machine Learning, Disease Prediction cardiac, K nearest neighbors, Genetic algorithm, Cleveland Heart Disease dataset.*

...

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

MNT	Maladies non transmissibles
MC	Maladie cardiaque
AVC	Accident vasculaire cérébral
ECG	électrocardiographie
IA	Intelligent artificielle
ML	Machine learning
CHDD	Cleveland Heart Disease dataset
KNN	K plus proches voisins
AG	Algorithme génétique
AG-KNN	Algorithme génétique et K plus proche voisins
LR	Logistique Régression
SVM	Support à vecteur machine
Rf	Random forest
DT	Decision tree
TP	True positive
TN	True négative
FP	False positive
FN	False négative
AED	Analyse exploratoire des données

Introduction générale

La santé humaine est une richesse, il n y'a rien de plus précieux qu'une bonne santé. Les chercheurs ont consacré de vastes efforts à proposer de nouvelles politiques, algorithmes, systèmes et architectures pour les soins de santé. Les soins de santé sont définis comme l'amélioration de la santé par la prévention, le traitement et l'examen des dommages de maladies et blessures. Le monde d'aujourd'hui est confronté à trois défis en matière de soins de santé : la pénurie de personnel médical, le vieillissement de la population et les dépenses élevées de soins de santé.

Des rapports de l'organisation mondiale de la santé (OMS) ont indiqué que les besoins mondiaux et le nombre factuel de personnels de santé étaient respectivement de 60,4 millions et 43 million en 2013. Ces chiffres passeront à 81,8 million et 67,3 million respectivement, d'ici 2030 [14]. Par conséquent, la pénurie de personnel médical n'est pas résolue et reste grave. De 2000 à 2050, le pourcentage de la population mondiale . de plus de 60 ans doublera (de 11 % à 22 %) Plus la personne est âgée, plus le risque de contracter des maladies et de nécessiter des soins et des traitements médicaux à long terme sont élevés.

Les maladies cardiaques sont parmi les maladies mortelles dans le monde, une grande proportion de gens souffre de ce problème. La détection et le traitement des maladies cardiaques (MCs) dans les premiers stades ont une grande importance. La bonne nouvelle est qu'avec l'augmentation rapide de la puissance de calcul et de la disponibilité des données de santé, les applications de santé peuvent inclure l'intelligence artificielle et devenir ainsi des soins de santé intelligentes. Des techniques d'identification des MCs par apprentissage automatique ont été développées pour aider les médecins et résoudre certains de défis susmentionnés en matière de santé.

L'apprentissage automatique est défini comme le domaine d'étude qui donne aux ordinateurs la capacité d'apprendre sans être explicitement programmés. De nombreux problèmes du monde réel peuvent être modélisés par des algorithmes d'apprentissage automatique comme la détection de visage, la reconnaissance vocale, la prédiction de maladies... etc.

L'objectif de ce travail est de proposer un système de prédiction de maladies cardiaques en utilisant l'algorithme d'apprentissage supervisé k-nearest neighbors (KNN). L'un des inconvénients majeurs de l'algorithme KNN est la vulnérabilité au bruit des attributs non pertinents et redondants. Ce bruit a des effets négatifs sur la performance du modèle prédictif. Divers chercheurs utilisent des techniques de minimisation de variables avant d'utiliser le KNN afin d'améliorer sa capacité prédictive et réduire le temps de calcul du système. Dans ce projet, on a utilisé l'algorithme génétique pour sélectionner les attributs pertinents et non redondants. Pour montrer la performance de notre modèle, on va le comparé avec d'autres approches d'apprentissage supervisées tel que Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), Suppor Vector Machine (SVM). Nous avons utilisé les données médicales « Cleveland Heart Disease dataset» pour l'entrainement et le test du modèle. Le reste du manuscrit est structuré comme suit :

— Le chapitre 1 s'intitule "Généralités sur les maladies cardiaques" présente brièvement des généralités sur les maladies cardiaques en détaillant la structure et le rôle du cœur humain, les causes et les facteurs de risque des maladies cardiaques, nous clôturons ce chapitre par quelques conseils pour prévenir les maladies cardiaques.
— Le deuxième chapitre " Apprentissage automatique et sélection d'attributs " est organisé en deux parties : Dans la première partie nous présentons un aperçu de l'apprentissage automatique avec une explication de l'algorithme KNN, dans la deuxième partie nous présentons le concept de sélection d'attributs et l'algorithme génétique.
— Dans le troisième chapitre " Analyse et approche proposée " : nous présentons la base de données sur laquelle nous avons travaillée et l'architecture proposée pour la résolution du problème de prédiction des maladies cardiaques.
— Enfin le dernier chapitre " Résultats et évaluation " : nous mettons en place les techniques et les outils que nous avons utilisés pour mettre en œuvre l'architec-

ture proposée et l'évaluation des résultats obtenues de l'approche proposée dans le chapitre 3.

Finalement nous conclurons notre travail avec une conclusion générale qui présente l'intérêt de l'approche hybride AGKNN pour la prédiction des maladies cardiaques et nous proposons quelques perspectives à faire dans des prochaines études dans ce domaine.

Chapitre 1

Généralités sur les maladies cardiaques

1.1 Introduction

Les maladies cardiaques sont parmi les principales causes de décès dans le monde : plus de personnes meurent chaque année de maladies cardiaques que de toute autre cause. C'est pour cela un diagnostic précoce et précis et un traitement de suivi approprié sauvent de nombreuses vies.

Dans ce chapitre, on va présenter des généralités sur les maladies cardiaques dont on va décrire en premier lieux la structure, le rôle et le fonctionnement du cœur humain, ensuite on présente les maladies cardiaques et leurs facteurs de risques afin de mieux comprendre notre sujet d'étude et enfin, on donne quelques conseils pour prévenir les maladies cardiaques.

1.2 Structure et rôle du cœur

Le coeur est un organe musculaire creux permet la circulation du sang dans le corps et l'apport d'oxygène et nutriments à l'ensemble des cellules des organismes. Le coeur est situé dans la partie médiane de la cage thoracique (le médiastin) délimité par les 2 poumons, le sternum et la colonne vertébrale. (voir figure 1.1)

Il est un peu plus gros chez l'homme que chez la femme et pompe chaque jour en moyenne 8000 litres de sang grâce à environ 100000 battements quotidiens (soit jusqu'à 2 milliards de battements au cours de la vie).

Le cœur est composé de 4 cavités : les oreillettes (ou atria) sur la partie supérieure et les

ventricules sur la partie inférieure. Les oreillettes et ventricules sont séparés de chaque côté par une épaisse paroi musculaire, le septum. Ainsi, aucun échange de sang entre la partie supérieur et la partie inférieure n'est possible. Le passage est unidirectionnel entre oreillette vers ventricule et cela via les valves cardiaques.

La paroi du cœur est composée de muscle en trois couches distinctes : l'épicarde (cellules épithéliales et tissu conjonctif), le myocarde ou muscle cardiaque et à l'intérieur, l'endocarde (cellules épithéliales et tissu conjonctif). Le sang qui circule dans le cœur va trop vite pour y être absorbé, si bien qu'il dispose de son propre système de vaisseaux, appelé artères coronaires, le vascularisant pour apporter aux cellules cardiaques oxygène et nutriments. Ce sont les ventricules qui assurent la fonction de pompes du sang vers le corps ou vers les poumons et leur parois est plus épaisse et leurs contraction plus fortes que les oreillettes [15].

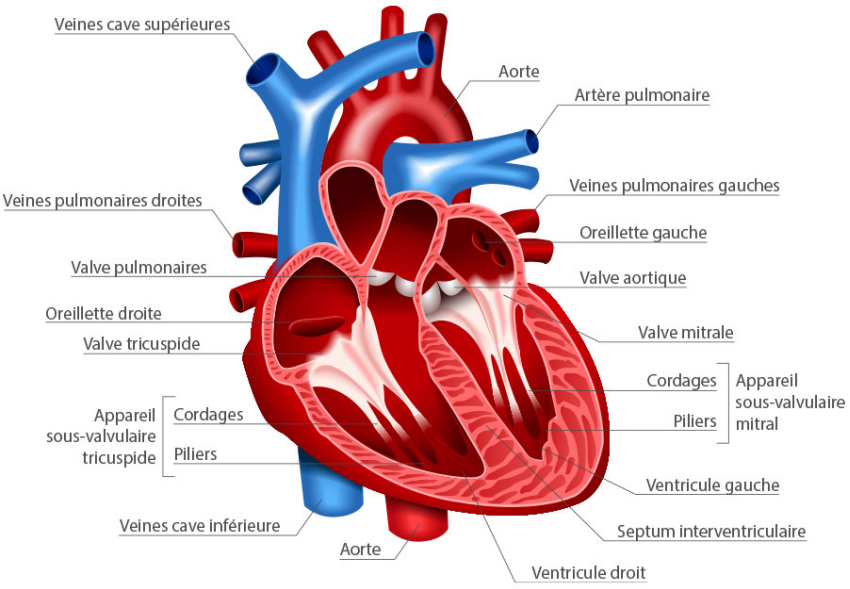

FIGURE 1.1 – Structure anatomique du cœur [1].

1.3 Fonctionnement du cœur

Le fonctionnement du coeur est shématisé par la figure1.2 :

1. Le côté droit du cœur renvoi le sang pauvre en oxygène aux poumons pour éliminer le dioxyde de carbone et réoxygéner le sang.

2. L'oreillette droite reçoit le sang veineux apporté par la veine cave et propulsé dans le ventricule droit qui en se contractant envoi le sang dans les poumons via l'artère pulmonaire (qui est donc la seule artère transportant du sang pauvre en oxygène).
3. Le sang oxygéné dans les poumons revient alors de cœur gauche au niveau de l'oreillette via les 4 veines pulmonaires (ce sont les seules veines transportant du sang riche en oxygène).
4. Le sang est ensuite propulsé dans le ventricule gauche et doit traverser la valve mitrale, qui contrôle le débit.
5. En se contractant, le cœur propulse via la valve aortique puis l'aorte (plus gros vaisseau sanguin de l'organisme) le sang dans l'ensemble du réseau des artères. Ce processus est répété 50 à 60 fois par minute au repos [15].

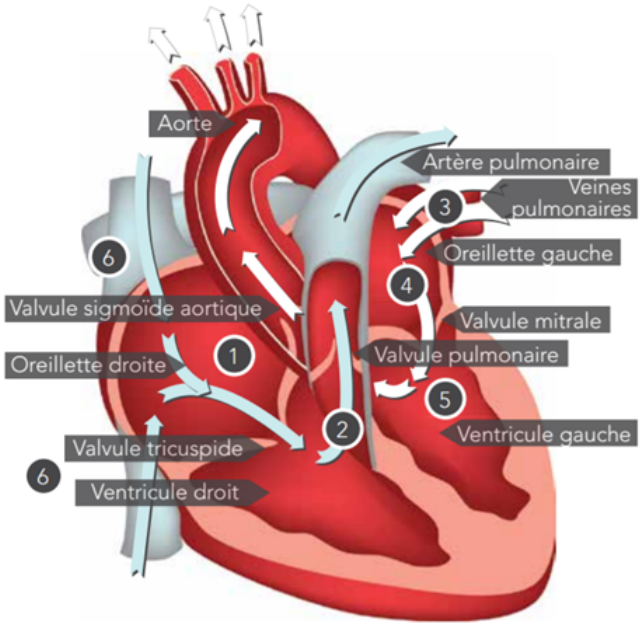

FIGURE 1.2 – Fonctionnement du cœur[2].

1.4 Les maladies cardiaques

Les maladies cardiques englobent plusieurs types de troubles de l'appareil circulatoire,la pathologie du coeur et les vaissaux sanguins, soit les maladies congénitales, ischémiques ou coronariennes, les maladies cérébrovasculaires et vasculaires périphériques ainsi que l'insuffisance et le rhumatisme cardiaque.

La maladie ischémique ou coronarienne est la maladie cardiaque la plus répandue. Elle touche les vaisseaux sanguins qui irriguent le muscle cardiaque. La maladie cérébrovasculaire est un problème au niveau de la circulation sanguine dans les vaisseaux du cerveau alors que la maladie vasculaire périphérique affecte principalement les vaisseaux qui alimentent les bras et les jambes.

L'insuffisance cardiaque survient lorsque le coeur ne pompe pas suffisamment de sang pour atteindre le niveau de circulation sanguine nécessaire aux besoins énergétiques du corps. Le rhumatisme cardiaque est une maladie infectieuse qui affecte les articulations et les valvules cardiaques alors que la maladie congénitale est une malformation du cœur qui découle d'une anomalie présente à la naissance [16].

— **Infarctus du myocarde (IDM)** L'infarctus du myocarde est une nécrose du myocarde se manifestant lorsqu'une ou plusieurs artères coronaires s'obstruent. De ce fait une partie du cœur n'est plus approvisionnée en sang et en oxygène [16].

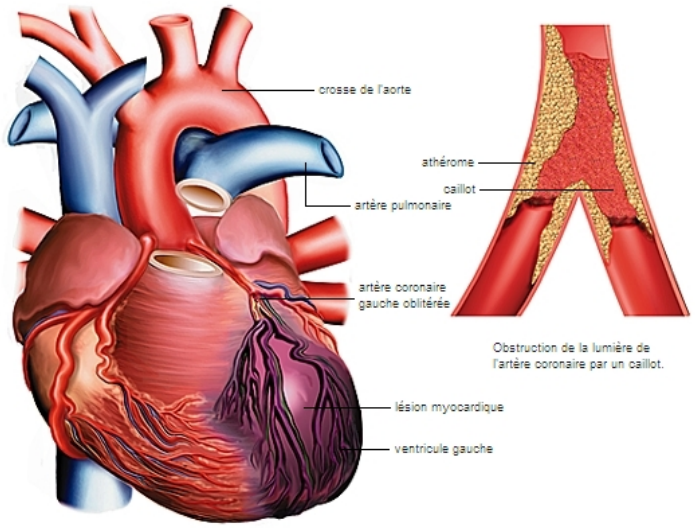

FIGURE 1.3 – Infarctus-du-myocarde [3].

— **Accident vasculaire cérébral (AVC)**

Un accident vasculaire cérébral est l'arrêt soudain de la circulation sanguine vers une ou plusieurs parties du cerveau. La rapidité de prise en charge est essentielle. Le symptôme le plus courant d'un accident vasculaire cérébral est une sensation soudaine de faiblesse au visage, au bras ou à la jambe, généralement d'un côté du corps.[16].

FIGURE 1.4 – Accident vasculaire cérébral (AVC) [4].

— **Athériosclérose**

L'athériosclérose est un épaississement et un durcissement de la paroi des artères [16].

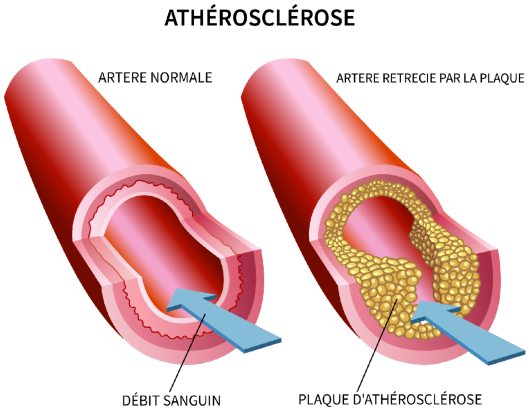

FIGURE 1.5 – Athérosclérose [5].

— **Angine de poitrine : Angor**

L'angine de poitrine d´écrit une douleur violente localisée dans la poitrine [16].

FIGURE 1.6 – Angine de poitrine [6].

— **Insuffisance cardiaque**

On parle d'insuffisance cardiaque lorsque le cœur n'est plus capable d'effectuer correctement son travail de pompe. Il est possible d'agir sur certains facteurs de risque (tabagisme, sur-poids, diabète...) en modifiant durablement ses habitudes de vie [16].

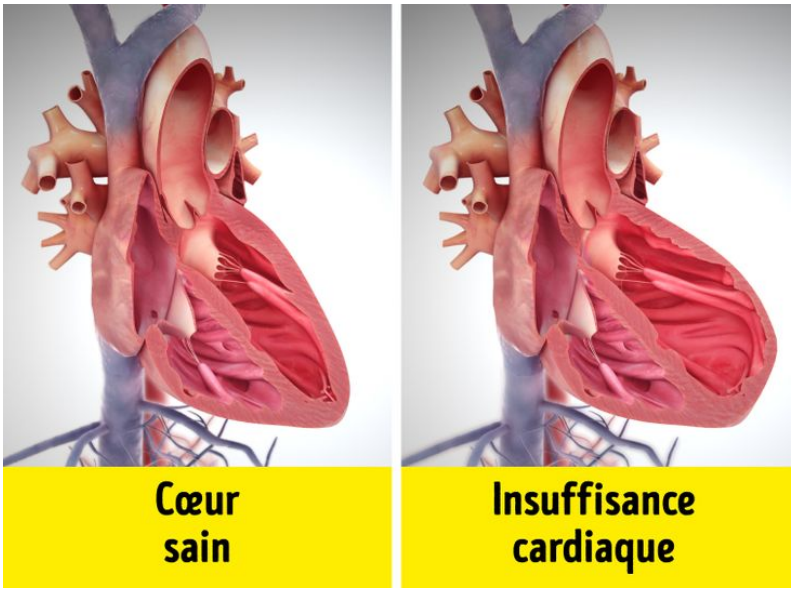

FIGURE 1.7 – La différence entre les deux cœurs [7].

— **Arythmie cardiaque**

L'arythmie cardiaque est un trouble du rythme cardiaque. Il en existe plusieurs types, de gravité variable [16].

— **Hypertension artérielle**

L'hypertension artérielle est une augmentation de la pression du sang dans les artères [16].

1.5 Diagnostique des maladies cardiaques

Si des symptômes évocateurs (fatigue, essoufflement, œdème, etc.) sont présents, une évaluation est nécessaire. Un examen physique permet aux médecins d'identifier l'insuffisance cardiaque, d'évaluer sa gravité et d'en trouver la cause (séquelles d'infarctus du myocarde, hypertension (HTA), valvulopathie, cardiomyopathie, etc.) [17] Les procédures diagnostiques peuvent être :

— Non invasives

— Minimalement invasives

Les tests non invasifs ne nécessitent pas d'incision ou de ponction à l'aiguille, parfois simplement en prélevant du sang ou en insérant un cathéter veineux court (IV) standard dans une veine du bras. Ces épreuves comprennent :

— Tomodensitométrie (TDM)
— Électrocardiographie (ECG)
— Imagerie par résonance magnétique (IRM)
— Tomographie par émission de positons (TEP)
— Scintigraphie
— Épreuve d'effort
— Test de la table basculante
— Échographie (y compris échocardiographie)
— Radiographies

Les tests mini-invasifs nécessitent généralement un long cathéter flexible qui est inséré dans un vaisseau sanguin du poignet, du cou ou de la cuisse, puis à travers le vaisseau sanguin jusqu'au cœur. Ces épreuves comprennent :

— Angiographie
— Cathétérisme cardiaque
— Cathétérisme veineux central
— Examen électrophysiologique

La plupart de ces procédures comportent peu ou pas de risque, mais le risque augmente avec la complexité de la procédure, la gravité de la maladie cardiaque et toute autre condition médicale dont souffre une personne. Le traitement peut parfois être administré lors de certains tests diagnostiques mini-invasifs. Par exemple, les patients atteints de maladie coronarienne peuvent subir une intervention coronarienne percutanée lors d'un cathétérisme cardiaque, tandis que les patients souffrant d'arythmies peuvent subir une ablation par radiofréquence lors d'examens électrophysiologiques.[18].

1.6 Les principaux symptômes des maladies cardiaques

Cette liste de symptômes n'est pas exhaustive, mais fournit des informations sur les symptômes les plus courants chez les personnes atteintes de maladies cardiaques. A l'inverse, la présence d'un de ces signes cliniques n'indique pas systématiquement la présence d'une cardiopathie. Les symptômes de chaque maladie sont sensiblement différents [19] :

— Douleur thoracique
— Essoufflement

— Fatigue
— Palpitations, perception des battements du cœur qui sont lents, rapides ou irréguliers
— Évanouissement, perte de connaissance, syncope
— Sensation de vertige
— Nausées et vomissements
— Douleur dans un membre, engourdissement ou crampe d'un muscle
— Gonflement des membres inférieurs (jambes, chevilles, pieds)
— Changement de la couleur de la peau au niveau d'un membre

1.7 Facteurs de risque des maladies cardiaques

Facteurs de risque non modifiables

— **Âge** : C'est un facteur de risque persistant qui augmente progressivement l'incidence de l'athérosclérose aortique, coronarienne et carotidienne et des complications de l'insuffisance cardiaque. Ce risque devient significatif à partir de 50 ans pour les hommes et 60 ans pour les femmes.
— **Sexe masculin** : Avant l'âge de 70 ans, les deux tiers des crises cardiaques surviennent chez les hommes. Cette différence diminue chez les femmes ménopausées et disparaît après 75 ans.
— **Hérédité** : Les antécédents familiaux de maladie cardiaque, coronarienne, d'accident vasculaire cérébral ou de mort subite sont des facteurs de risque, surtout s'ils surviennent chez un ou plusieurs parents du premier degré, à un âge jeune (< 55 ans pour le père ou < 65 ans pour la mère) [20].

Facteurs de risque modifiables

— **Tabagisme :**

Il augmente les lésions athérosclérotiques en modifiant la fonction endothéliale, en perturbant la vasomotion, en activant l'agrégation plaquettaire et en réduisant le cholestérol des lipoprotéines de haute densité. Elle est athérogène et prothrombotique. Le risque relatif de crise cardiaque était de 5, et le risque relatif de maladie

artérielle des membres inférieurs était de 2. Ce risque relatif existe également lors du tabagisme passif. Le risque est proportionnel à l'exposition au tabac et s'évalue en paquets-années. Les bénéfices de l'arrêt du tabac sont rapides : l'augmentation du risque relatif disparaît en 3 ans, et le risque de récidive est réduit de 50% chez les patients coronariens[20].

— **Hypertension artérielle :**

Elle est définie par une valeur de pression de $>$ 140 mmHg pour la pression artérielle systolique (PAS) ou de $>$ 90 mmHg pour la pression artérielle diastolique (PAD). Tous les types d'hypertension sont des facteurs de risque : hypertension permanente, hypertension paroxystique, traitée ou non.

Le risque relatif était de 7 points pour les accidents vasculaires cérébraux, 3 points pour les maladies coronariennes et 2 points pour les maladies artérielles des membres inférieurs. Avant 55 ans, ce risque était associé aux valeurs de tension artérielle systolique et diastolique. Après 60 ans, la corrélation est plus forte avec la pression différentielle (PAS - PAD), et donc la pression artérielle systolique surtout chez les personnes âgées. Le traitement de l'hypertension artérielle réduit le risque d'AVC de 40 % et le risque d'infarctus de 15 % [20].

— **Dyslipidémies :**

Parmi les dyslipidémies circulantes, le principal facteur de risque de maladie cardiovasculaire est l'élévation du cholestérol LDL, avec un cholestérol lié au LDL 1,60 g/L (4,1 mmol/L). Le cholestérol LDL était positivement associé au risque de maladie cardiovasculaire, tandis que le cholestérol HDL était associé négativement s'il était $>$ 0,40 g/L (1 mmol/L). Des taux élevés de triglycérides ($>$ 2,0 g/L) seuls ne sont pas un facteur de risque (indépendant), mais peuvent être un facteur de risque en combinaison avec d'autres facteurs (voir Syndrome métabolique). Le cholestérol à lipoprotéines de basse densité a un rôle direct dans la croissance des plaques d'athérosclérose et leur rupture due à l'instabilité. Le risque relatif d'hypercholestérolémie pour les maladies coronariennes était 3 fois plus élevé que pour les maladies artérielles et les accidents vasculaires cérébraux.

L'efficacité du traitement de l'hypercholestérolémie est un facteur majeur de réduction de la mortalité cardiovasculaire (30% à 20 ans)[20].

— **Diabète :**

Le diabète a été défini comme deux mesures à jeun > 1,26 g/L (7 mmol/L) ou une seule mesure de glycémie > 2 g/L (11 mmol/L). Le diabète de type I ou de type II est associé à un risque cardiovasculaire accru. Les complications cardiovasculaires du diabète I commencent plus tôt dans la trentaine, mais l'apparition rapide du diabète II en fait un facteur de risque très préoccupant.

Il a un risque relatif de >2, causant principalement plus de maladies artérielles que de maladies coronariennes et d'accidents vasculaires cérébraux. Mais le diabète se complique plus souvent de microangiopathie (rétinopathie et maladie rénale). Ce risque relatif augmente avec les anomalies rénales. Le traitement du diabète avec un objectif de 6,5 % d'hémoglobine glyquée (HbA1c) réduit les complications cardiovasculaires [20].

— **Insuffisance rénale**

L'insuffisance rénale chronique est associée à un taux élevé de complications cardiovasculaires, comparable à la sévérité du diabète sur le système cardiovasculaire [20].

1.8 Prévention des maladies cardiaques

Prévention individuelle

L'objectif est que chacun élimine ou réduise au maximum les facteurs de risque modifiables. Cela comprend l'élimination du tabac, le respect du mode de vie et des mesures diététiques et la pratique d'une activité physique régulière. Des médicaments peuvent également être prescrits en respectant les conseils selon [21] :

— Le résultat de ces premières mesures ;
— Le nombre de facteurs de risque ;
— Le calcul du risque absolu ;
— La présence d'antécédents cardiovasculaires.

Prévention collective

Son champ d'application est considérable, car il touche toute la population et comprend notamment [21] :

— La réglementation anti-tabac ;
— La limitation réglementée de la teneur en sel dans l'industrie alimentaire ;
— L'éducation et l'alimentation scolaires, principaux axes pour freiner l'obésité croissante ;
— L'information de la population par diverses campagnes nationales ou régionales ;
— L'accès aux équipements sportifs et zones de plein air.

1.9 Conclusion

Dans ce chapitre nous avons présenté de manière globale le cœur dans l'organisme humain, les maladies cardiaques : facteurs de risques et quelques conseils pour se prévenir. Dans le chapitre suivant nous allons s'intéresser au domaine de l'apprentissage automatique avec ces différentes étapes et le processus de sélection des attributs.

Chapitre 2

Apprentissage automatique & Sélection des attributs

2.1 Introduction

L'apprentissage automatique est l'un des domaines de recherche de l'intelligence artificielle. Il fait référence au développement, à l'analyse et à la mise en œuvre des modèles qui permettent aux machines d'évoluer pour effectuer des tâches difficiles ou impossibles à accomplir par des moyens algorithmiques traditionnels. Lors de la création de ces modèles sur un ensemble de données contenant des variables non importantes la précision globale du classificateur se réduit et sa complexité s'augmente, c'est pour cela nous utilisons les techniques de sélection d'attributs. Ce chapitre est organisé en deux parties : Dans la première partie nous allons présenter le domaine de l'apprentissage automatique, ses différents types et les algorithmes utilisés pour la réalisation de ce travail, Dans la deuxième partie nous présentons le concept de sélection des attributs et l'algorithme génétique.

2.2 Apprentissage automatique

L'apprentissage automatique ou machine learning (ML) en anglais est une application de l'intelligence artificielle (IA) (voir figure 2.1) qui permet aux systèmes d'apprendre et de s'améliorer automatiquement à partir de l'expérience sans être explicitement programmé. Il se concentre sur le développement des programmes informatiques capables d'accéder aux données et de les utiliser pour apprendre par eux-mêmes [22].

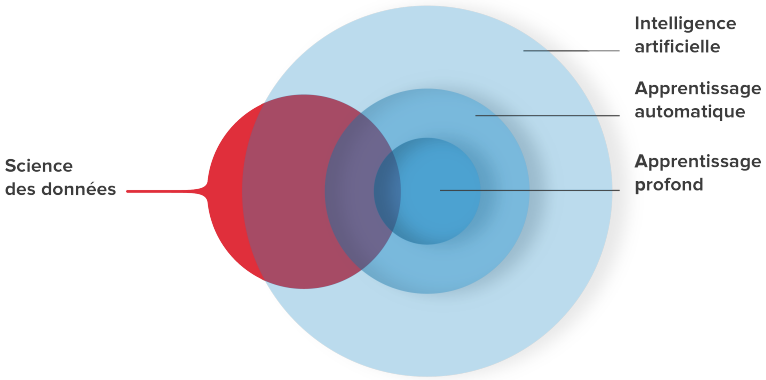

FIGURE 2.1 – Relation IA, Apprentissage Automatique et Apprentissage Profond [8].

2.3 Types d'apprentissage automatique

L'apprentissage automatique implique différents types d'apprentissage qui peuvent se catégoriser selon la structure du problème qu'ils emploient. La figure 2.2 illustre les différents types de l'apprentissage automatique les plus courant dans les principaux domaines d'application : Apprentissage supervisé, non supervisé et par renforcement.

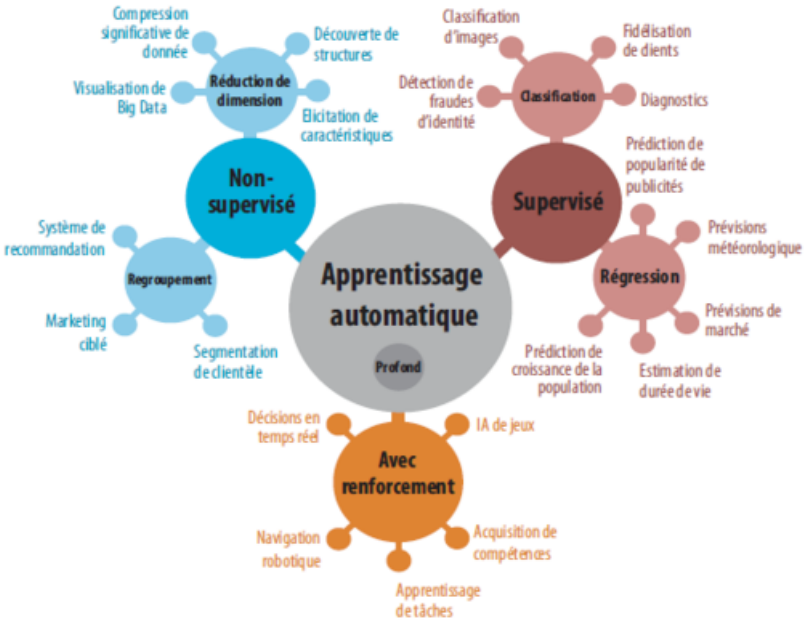

FIGURE 2.2 – Classification des différents types d'apprentissage automatique [9].

2.3.1 Apprentissage supervisé

L'apprentissage automatique supervisé exige à ses algorithmes d'utiliser des exemples étiquetés pour appliquer ce qu'ils ont appris dans le passé aux nouvelles données afin de

prédire des événements futurs à partir de l'analyse d'un ensemble de données déjà connues, dont ils produit une fonction déduite pour faire des prédictions sur la valeur de sortie, il peut également comparer sa sortie aux sorties correctes et prédites et trouver des erreurs pour modifier le modèle en conséquence.

En apprentissage supervisé, on distingue entre deux types de tâches [23] :

- **La classification :** Quand la variable cible (à prédire) est discrète. Ce qui revient à attribuer une classe (ou étiquette) à chaque entrée. C'est le cas si on cherche à prédire la tendance d'un mouvement futur d'un actif (haut, neutre, bas).
- **La régression :** Quand la variable cible à prédire est continue.

Voici quelques algorithmes d'apprentissage supervisé les plus importants [24] :

- K plus proches voisins ;
- Régression linéaire ;
- Régression logistique ;
- Machines à vecteurs de support (SVM) ;
- Arbres de décision et forêts aléatoires.

2.3.2 Apprentissage non supervisé

Pour ce type d'apprentissage la base de données d'apprentissage ne contient pas de variable cible (comme on l'a vu en apprentissage supervisé). Il y a seulement un ensemble de données collectées en entrée. L'algorithme doit découvrir par lui-même la structure en fonction des données [23].

Nous citons quelques algorithmes d'apprentissage non supervisé les plus importants [24] :

- Clustering :
  - K-Means ;
  - Analyse des clusters hiérarchiques (HCA) ;
  - Maximisation des attentes.
- Visualisation et réduction de la dimensionnalité :
  - Analyse en composantes principales (ACP) ;
  - Kernel PCA ;
  - L'encastrement linéaire local (LLE) ;
  - T-distribué Stochastic Neighbor Embedding (t-SNE).
- Apprentissage des règles d'association :

— Apriori ;

— Eclat.

2.3.3 Apprentissage par renforcement

L'apprentissage se fait sans supervision, par interaction avec l'environnement (principe d'essai / erreur) et en observant le résultat des actions prises. Chaque action de la séquence est associée à une récompense. Le but est de déterminer la stratégie comportementale optimale afin de maximiser la récompense totale. Pour cela, un simple retour des résultats est nécessaire pour apprendre comment la machine doit agir. Ceci est appelé le signal de renforcement. Il peut être très avantageux pour la prévision financière à haute fréquence où l'environnement est dynamique et en conséquence, il est difficile de trouver ou d'automatiser manuellement des stratégies efficaces [24].

2.4 Algorithmes d'apprentissage automatique supervisés utilisés

Dans cette section nous représenterons les algorithmes d'apprentissage automatique utilisés dans notre étude.

2.4.1 Régression logistique

La régression logistique ou logistique régression (LR) en anglais est un modèle statistique permettant d'étudier les relations entre un ensemble de variables qualitatives Xi et une variable qualitative Y. Un modèle de régression logistique permet aussi de prédire la probabilité qu'un événement arrive (valeur de 1) ou non (valeur de 0) à partir de l'optimisation des coefficients de régression. Ce résultat varie toujours entre 0 et 1. Lorsque la valeur prédite est supérieure à un seuil, l'événement est susceptible de se produire, alors que lorsque cette valeur est inférieure au même seuil, il ne l'est pas [25].

2.4.2 Machines à vecteurs de support

Les machines à vecteurs de support ou support vector machine (SVM) en anglais, sont des modèles de l'apprentissage automatique supervisé centrés sur la résolution des

problèmes de discrimination et de régression mathématiques. Il consiste à ramener un problème de classification ou de discrimination à un hyperplan (feature space) dans lequel les données sont séparées en plusieurs classes dont la frontière est la plus éloignée possible des points de données (ou "marge maximale"). D'où l'autre nom attribué aux SVM : les séparateurs à vaste marge. Le concept de frontière implique que les données soient linéairement séparables. Pour parvenir, les SVM font appel à des fonctions mathématiques permettant de projeter et séparer les données dans l'espace vectoriel, les "vecteurs de support" étant les données les plus proches de la frontière. C'est la frontière la plus éloignée de tous les points d'entraînement qui est optimale, et qui présente donc la meilleure capacité de généralisation [26].

2.4.3 Arbres de décision

Les arbres de décision ou decision tree (DT) en anglais sont des algorithmes de prédiction qui fonctionnent en régression et en classification. Ils permettent de trouver une partition qui sépare le plus possible les différentes observations. Après la segmentation, un ensemble de règles est créé pour prédire un résultat ou une classe.
En théorie des graphes, un arbre est un graphe non orienté, acyclique et connexe .
L'ensemble de nœud est divisé en trois catégories [27] :

1. **Nœud racine** : Ce nœud est utilisé pour accéder à l'arborescence.
2. **Nœuds internes** : nœuds qui ont des descendants.
3. **Nœuds d'extrémité(ou feuilles)** : Nœuds qui n'ont pas de descendants.

Chaque individu auquel une classe doit être attribuée est décrit par un ensemble de variables testés dans les nœuds de l'arbre. Les tests sont effectués dans les nœuds internes et les décisions sont prises dans les nœuds feuilles .

2.4.4 Forêts aléatoires

Foret aléatoire ou Random forest (RF) en anglais est une technique célèbre de l'apprentissage automatique, Le nom de cette technique "RandomForest" nous indique qu'elle est basée sur un ensemble d'arbres que l'on appelle les arbres de décision, ces derniers sont des outils qui aident à la décision, permet de représenter un ensemble de choix sous la forme graphique d'un arbre, elles sont interprétable et facile à utiliser et à entraîner mais

elles ne sont pas généralisables. C'est une solution pour régler le problème de généralisation en assemblant plusieurs arbres de décision pour construire une foret, cette technique va produire des résultats généralisable. On peut résumer les étapes de l'algorithme de Random Forest en trois étapes [28] :

1. La création des n arbres de décision à partir des nouvelles matrices de données.
2. Faire entraîner chaque arbre de décision sur une nouvelle matrice de données.
3. Prendre une décision majoritaire parmi les décisions obtenues par les n arbres.

Pour ce projet on va s'intéresser à l'algorithme d'apprentissage supervisé K plus proches voisins que nous allons le détailler dans la sections suivante.

2.4.5 K plus proches voisins

L'algorithme des k les plus proches voisins ou k-nearest neighbors (KNN) en anglais est un algorithme de classification supervisée consiste à affecter une classe à un vecteur de paramètres de test en comparant ce dernier à un ensemble de vecteurs étiquetés, préalablement enregistré durant la phase d'apprentissage. Cette comparaison vise à sortir parmi cet ensemble les k vecteurs les proches au vecteur considéré, en termes de distances. La classe affectée au vecteur de test est la classe la plus votée parmi les k classes obtenues dans l'étape de comparaison [29].
La valeur de K et le choix de la distance sont les paramètres les plus importants de l'algorithme de classification KNN.

- Le choix de la meilleur valeur de K : En règle générale, le meilleur choix de k dépend des données. La valeur optimale de k peut être choisie par des techniques heuristiques [30]. Généralement, le choix de la valeur optimale de K consiste à chercher parmi un ensemble de valeurs K, celle qui permet d'obtenir un taux de classification maximal évalué sur une base de données de test.
- Le choix de la distance optimale : Il consiste à suivre la même stratégie du choix de la valeur de K en appliquant différentes distances utilisées dans l'algorithme KNN. Les distances utilisées sont décrites comme suit [31] :
  - **La distance Euclidienne :** Cette distance a été utilisée dans plusieurs systèmes d'identification basée sur l'algorithme KNN [32]

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2.1}$$

Où : x, y sont des vecteurs.

— **La distance de Minkowsky [32] :**

$$d(x,y) = \sqrt[p]{\left(\sum_{i=1}^{n} |x_i - y_i|^p\right)} \tag{2.2}$$

Où : x, y sont des vecteurs. p : paramètre

— **La distance de Manhattan [32] :**

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i| \tag{2.3}$$

Où : x, y sont des vecteurs

Les étapes de l'algorithme KNN sont décrits par le pseudo code suivant[33] :

L'algorithme des k plus proches voisins

---

1. Début Algorithme

2. Données (ou vecteurs) en entrée

- un ensemble de données (base de données).
- une fonction pour la définition de la distance.
- Un nombre entier K.

3. Pour une nouvelle observation (vecteur) dont on veut prédire sa classe Faire :

4. Calculer toutes les distances de cette observation avec les autres observations du jeu de données (base de données d'apprentissage).

5. Retenir les observations du jeu de données les proches en utilisant la fonction de calcul de distance.

6. Prendre les valeurs des classes d'observations retenues. Retourner la classe la plus dominante ou la plus votée.

7. Fin Algorithme

2.4.5.1 Exemple illustratif de classification par KNN

Un exemple de classification KNN est illustré dans la figure 2.3 tel que :
le point inconnue (*étoile*) appartient soit à la première classe (*carré*)ou à la deuxième classe (*triangle*).
Si K =3, le point inconnue est classé dans la deuxième classe (*Triangle*) parce qu'il y a 2 triangles et un seul *carré* parmi les trois plus proches exemple a l'intérieur du *cercle* .
Si K = 5, il est classé dans la première classe *carré* .

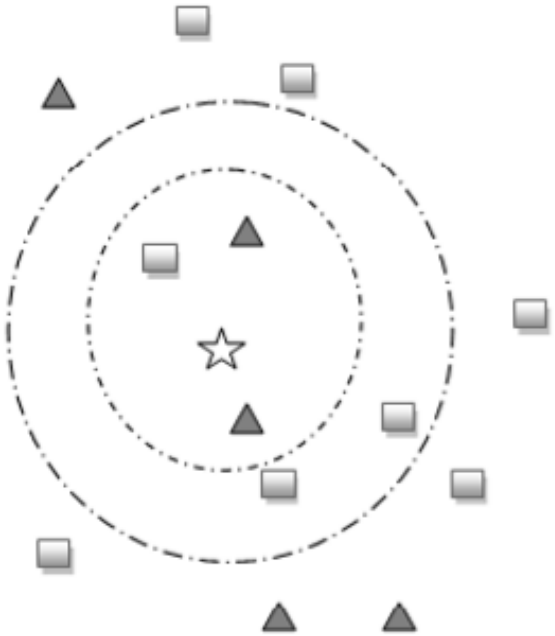

FIGURE 2.3 – Exemple de classification KNN (K=3 et K = 5) [10].

2.4.5.2 Avantages et inconvénients de KNN

La méthode des k plus proches voisins représente des avantages ainsi des inconvénients tels que :

2.4.5.3 Les avantages de la méthode des k plus proches voisins

1. L'algorithme KNN est sensible envers des données bruitées[34].

2. La méthode des k plus proches voisins est efficace si les données sont larges et incomplètes[35].
3. Cette méthode est l'une des plus simples de tous les algorithmes d'apprentissage automatique[32].

2.4.5.4 Inconvénients de la méthode des k plus proches voisins

1. Le besoin de déterminer la valeur du nombre des plus proches voisins (le paramètre **k**)[36].
2. Le temps de prédiction est très long puisqu'on doit calculer la distance de tous les exemples[34].
3. Cette méthode est gourmande en espace mémoire car elle utilise une grande capacité de stockage pour le traitement des corpus[37].

2.5 Sélection des attributs

La sélection d'attributs ou feature selection (FS) en anglais est un processus de réduction de dimensionnalité qui consiste à choisir à partir d'un ensemble d'attributs un sous ensemble d'attributs pertinents en supprimant les attributs non appropriés, redondantes ou bruyantes, Dans l'objectif d'avoir de meilleures performances d'apprentissage, c'est-à-dire une plus grande précision d'apprentissage, un minimum coût de calcul et une meilleure interprétabilité du modèle [38]. Ce processus sélectionne les attributs en se basant sur certains critères pour éliminer tous les facteurs qui ne sont pas en rapport avec le problème traité, et garder efficacement les attributs importants, l'importance de ces attributs est classée selon leurs notion de pertinence comme suite [39] :

— **Les variables fortement pertinentes** : Ils sont donc indispensables et devraient figurer dans tout sous-ensemble optimal sélectionné, car leurs absences peuvent conduire à un défaut de reconnaissance de la fonction cible (la classe).
— **Les faibles pertinentes** : Elle suggère que la variable n'est pas toujours importante, mais il peut devenir nécessaire pour un sous-ensemble optimal dans certaines conditions.
— **Les non-pertinentes** : La non-pertinence d'une variable se définit simplement et indique qu'une variable n'est pas du tout nécessaire dans un sous-ensemble optimal

de variables.

2.5.1 Processus de sélection des attributs

La figure 2.4 illustre le processus de sélection d'attribut. Dont on distingue 4 étapes, en commençant par l'ensemble initial des attributs, la génération du sous-ensemble, l'évaluation du sous ensemble, critère d'arrêt et la validation des résultats [38].

1. **La génération du sous ensemble :** est utilisé pour déterminer des sous-ensembles d'attributs candidats pour l'évaluation.
2. **L'évaluation du sous-ensemble :** est utilisé pour mesurer la qualité du sous-ensemble candidat. Pour qu'il se compare avec le meilleur sous-ensemble précédent pour déterminer si ce sous-ensemble est convenable ou non. Si le nouveau sous-ensemble candidat est meilleur, il remplace le précédent meilleur. En répétant ce processus jusqu'à atteindre la meilleure valeur d'évaluation.
3. **Critère d'arrêt :** il est nécessaire que chaque sous-ensemble d'attributs après l'évaluation soit comparé au critère d'arrêt pour vérifier si les attributs du sous-ensemble actuel ont atteint un niveau prédéfini. Si les exigences sont vérifiés, la sélection d'attributs s'arrête et le sous-ensemble courant est considéré comme le résultat final ; sinon le processus de recherche continue.
4. **La validation :** est faite par différents test avec des données du monde réel ou non réel.

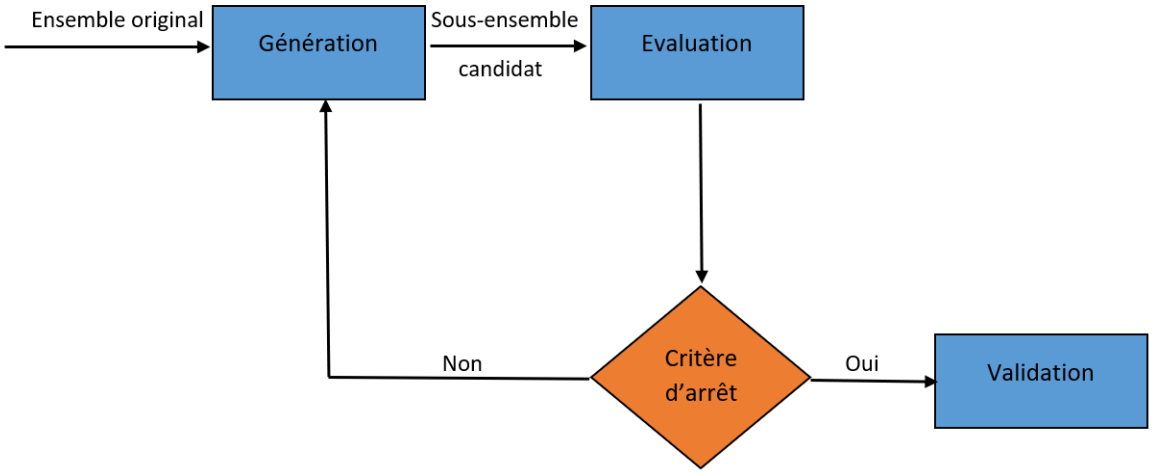

FIGURE 2.4 – Processus de sélection des attributs.

2.5.2 Les algorithmes génétiques

Les algorithmes génétiques (AGs) sont des algorithmes de recherche basés sur les mécanismes de la sélection naturelle et de la génétique. Il combine une stratégie de "survie des plus forts" avec un échange d'information aléatoire mais structuré [40], Les AGs utilisent un vocabulaire similaire à celui de la génétique naturelle, nous retrouvons les notions de Population, d'Individu, de Chromosome et de Gène (voir figure 2.5 [41, 42].

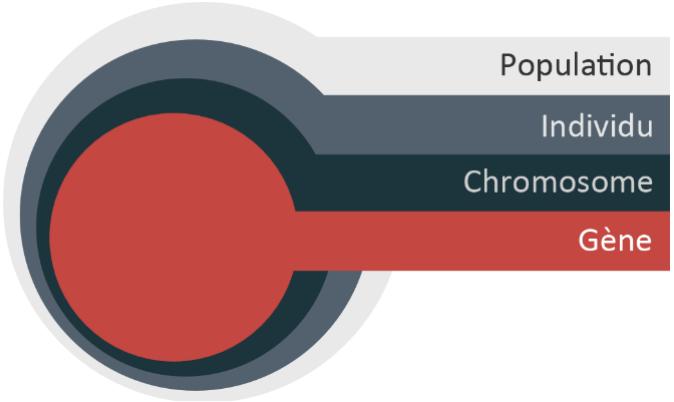

FIGURE 2.5 – Paradigme de terminologie de l'algorithme génétique [11].

— **La population** est l'ensemble des solutions envisageables.

— **L'individu** représente une solution.

— **Le Chromosome** est une composante de la solution.

— **Le Gène** est une caractéristique, une particularité.

Avec ces notions, nous obtenons trois opérateurs d'évolution génétiques :

— **La sélection :** Choix des individus les mieux adaptés.

— **Le croisement :** Mélange par la reproduction des particularités des individus choisis.

— **La mutation :** Altération aléatoire des particularités d'un individu.

2.5.2.1 Principe de fonctionnement de l'algorithme génétique

Le principe général du fonctionnement d'un algorithme génétique est représenté sur la figure 2.6,on commence par la génération de la population initiale, une fonction d'évaluation qui dépend de la fonction à optimiser, des opérateurs permettant de diversifier la population au cours des générations et d'explorer l'espace de recherche (sélection, croisement, mutation). Il faut également choisir les paramètres de l'algorithme qui sont nombreux et parfois délicats à régler : probabilité de croisement P_c et de mutation P_m , taille de la population et nombre de générations (si c'est le critère d'arrêt choisi) [12].

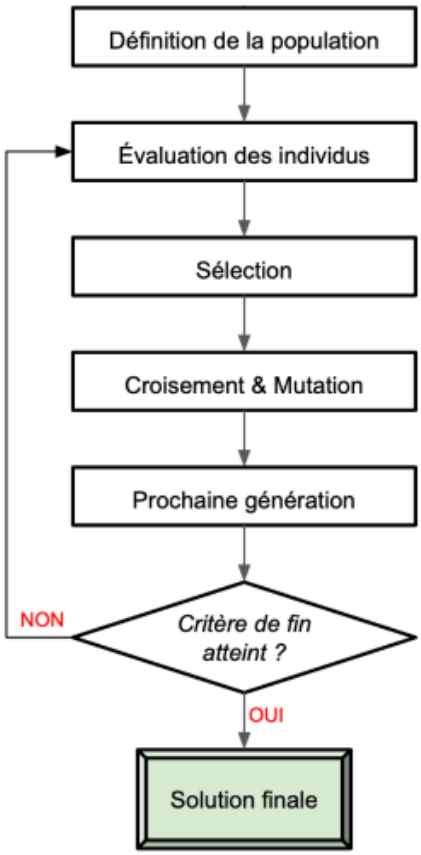

FIGURE 2.6 – Principe du fonctionnement d'un algorithme génétique [12].

2.6 Évaluation des performances des modèles de classification

L'évaluation permet de tester le modèle par rapport à des données qui n'ont jamais été utilisées pour l'entraînement. Cela permet de voir comment le modèle pourrait fonctionner par rapport à des données qu'il n'a pas encore vues. Ceci est censé être représentatif de la façon dont le modèle pourrait fonctionner dans le monde réel.
Pour évaluer les performances d'un modèle de classification on utilise la matrice de confusion.

2.6.1 Matrice de confusion

Une matrice de confusion est un tableau qui résume des prédictions pour un problème de classification particulier. Il compare les données réelles de la variable cible avec les données prédites par le modèle [43]. Les prédictions correctes et fausses sont affichées et réparties par 4 catégories (voir la figure 2.7) expliquées comme suite :

1. **True Positive (TP)** : la prédiction et la valeur réelle sont positives.
   Exemple : Une personne malade et prévu malade.
2. **True Negative (TN)** : la prédiction et la valeur réelle sont négatives.
   Exemple : Une personne saine et prévu saine.
3. **False Positive (FP)** : la prédiction est positive alors que la valeur réelle est négative.
   Exemple : Une personne saine et prévu malade.
4. **False Negative (FN)** : la prédiction est négative alors que la valeur réelle est négative.
   Exemple : Une personne malade et prévu saine.

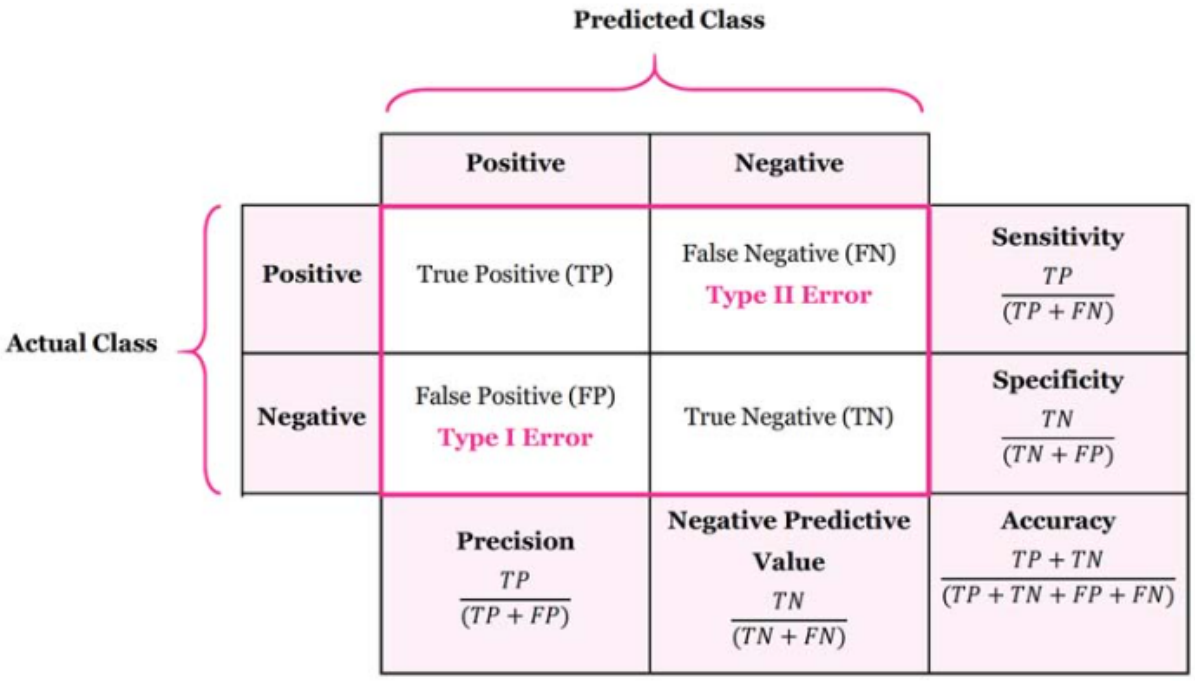

FIGURE 2.7 – Matrice de confusion [13].

La matrice de confusion est utilisée pour calculer les métriques de performences : Accuracy, Rappel, F1-score, Précision [43].

2.6.1.1 Accuracy

L'exactitude permet de connaître la proportion de bonnes prédictions par rapport à toutes les prédictions. L'opération est simplement : Nombre de bonnes prédictions / Nombre total de prédictions [43].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.4}$$

2.6.1.2 Rappel

Le rappel correspond au nombre de documents correctement attribués à la classe i par rapport au nombre total de documents appartenant à la classe i (total true positive) [43].

$$Recall = \frac{TP}{TP + FN} \tag{2.5}$$

2.6.1.3 Précision

La précision correspond au nombre de documents correctement attribués à la classe i par rapport au nombre total de documents prédits comme appartenant à la classe i (total predicted positive) [43].

$$Precision = \frac{TP}{TP + FP} \tag{2.6}$$

2.6.1.4 F1score

Le $F1 - score$ est une métrique pour évaluer la performance des modèles de classification à 2 classes ou plus. Il est particulièrement utilisé pour les problèmes utilisant des données déséquilibrées [43].
Le $F1 - score$ permet de résumer les valeurs de la précision et du recall en une seule métrique. Mathématiquement, le F1-score est défini comme étant la moyenne harmonique de la précision et du recall, ce qui se traduit par l'équation suivante :

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{2.7}$$

2.7 Conclusion

Nous avons présenté à travers ce chapitre, de façon générale, l'apprentissage automatique avec ses différents types, comme nous avons cité les méthodes de la sélection des attributs en présentant l'algorithme génétique.
A la fin, nous avons vu les techniques qui peuvent être utiliser pour l'évaluation des performances des modèles d'apprentissage automatique .
Le chapitre qui suit sera l'application de contenu de ce chapitre sur notre étude .

Chapitre 3

Analyses et approche proposée

3.1 Introduction

La maladie cardiaque est une maladie très mortelle. Dans le monde, une grande proportion de gens souffre de ce problème. La détection des maladies cardiaques (MC) à l'aide de modèles d'apprentissage automatique est très efficace dans les premiers stades. Le traitement de MC est efficace s'il détecte la maladie aux stades initiaux. Des techniques d'identification de MC par apprentissage automatique ont été développées pour aider les médecins. Dans ce chapitre, on va proposer un système d'identification en utilisant un modèle de ML basé sur le KNN pour classer les malades cardiaques et les personnes sains. L'algorithme génétique est utilisé pour sélectionner les attributs les plus appropriés afin d'augmenter la précision de la classification et de réduire le temps de calcul du système prédictif. Pour montrer la performance de notre modèle, on va le comparé avec d'autres approches d'apprentissage automatique tel que RF, DT, LR, SVM.

3.2 Etat de l'art

Au cours des dernières années de diverses recherches ont soulignés le potentiel de classification dans la prédiction des maladies cardiaques
Dans cette section, nous expliquons certaines des solutions récentes, le tableau 3.1 résume les résultats des solutions étudiées :

— V. Sharma, A. Rasool, and G. Hajela [44] ont fait des recherches sur la prédiction des maladies cardiaques.avecla base de données de cleveland, Ces recherches ont

souligné que le réseau de neurones profonds (DNN) possède différentes propriétés telles que l'optimisation, la technique de pondération et le nombre de couches cachées. DNN a été exécuté avec trois autres algorithmes tels que Support Vector Machine (SVM), Naive Bayes (NB) et KNN. les résultats montrent que SVM a donné les meilleures performances que les autres.

— Md. Nahiduzzaman Md. Julker Nayeem [45] ont proposé deux classificateurs pour prédire les maladies cardiaques en utlisant lensemble de données de cleveland. L'un est un Perceptron multicouche réseau de neurones (MLP) et un autre est le vecteur de support Machine (SVM). leurs travail consiste à classer la maladie en deux classes (malade ou non malade) et cinq autres classes de maladies cardiaques.ils ont utilisé la Base de données en ligne sur les maladies cardiaques de Cleveland qui comprend 303 instances avec 5 classes et 13 attributs. dans le classement en cinq classes ils ont eu pour MLP une précision de 68,86% alors que SVM etait 59,01 %.

— Halima EL HAMDAOUI1, Saïd BOUJRAF [46] ont proposé un système de soutien clinique pour prédire les maladies cardiaques aider les cliniciens à diagnostiquer et à prendre de meilleures décisions. ils ont utilisé dans cette étude les algorithmes d'apprentissage automatique tels que Naïve Bayes, K plus proche Voisin, Support Vector Machine, Random Forest et L'arbre de décision en utilisant des données sur les facteurs de risque extraites des dossiers médicaux. le résultat révèle que Naïve Bayes surpasse en utilisant la validation croisée et la division train-test techniques avec une précision de 82,17%, 84,28%, respectivement.

— P. Sujatha and K. Mahalakshmi [47] dans leurs expériences pour prédire la maladie cardiaque en utilisant l'arbre de décision (DT), le NB, la forêt aléatoire (RF), le SVM, le KNN et la régression logistique (LR), ont déclaré que la forêt aléatoire est plus importante qu'aux autres et peut gérer un ensemble de données de petite taille. Ils ont suggéré d'effectuer différentes techniques d'apprentissage automatique pour la prédiction des maladies cardiaques avec un ensemble de données de grande taille afin de créer un modèle de prédiction fiable et précis en temps réel.

— Singh and R. Kumar [48] afin de déterminer le meilleur algorithme de prédiction des maladies cardiaques, ils ont comparé la précision de quatre algorithmes d'apprentissage automatique différents tels que l'arbre de décision, la régression logistique, le KNN et le SVM. Les expériences montrent que KNN donne le meilleur résultat.

— Revati et al. [49], ont fournit une analyse de diverses méthodes d'exploration de données pour la prédiction de maladie cardiaque en utilisant L'arbre de décision, l'algorithme de rétropropagation et Naive Bayes. Le système a utilisé 14 paramètres, dont la pression artérielle, les douleurs thoraciques, le cholestérol et la fréquence cardiaque, pour améliorer la précision du système. l'étude montre que le réseau de neurones fonctionne le mieux pour la prédiction avec une précision de 100 %. Il a surpassé les autres deux algorithmes.

— Pour améliorer l'efficacité du système de prédiction des maladies cardiaques, R. Sateesh Kumar et S. Sameen Fatima [50] ont utilisé un test du chi carré pour sélectionner les caractéristiques les plus importantes. Basé sur l'accuracy, le rappel, la précision et le score F1. Les résultats de l'algorithme proposé sont plus précis que tous les attributs avec moins d'attributs. La performance de E-KNN utilisant 11 attributs a une valeur d'accuracy de 90,10 %. suivi de svm avec une accuracy de 89%.

Références	**Date**	**Algorithmes**	**Accuracy**
V. Sharma, A. Rasool [44]	2020	- Naive Bayes	81,4%
		- SVM	86,20%
Md. Toukir Ahmed [45]	2019	- MLP	90,57%
		- SVM	92,57%
Jessica I. Gupta ,Michael [46]	2020	- Naive Bayes	84%
		- KNN	81,31%
		- SVM	81,42%
		- RF	77,14%
		- DT	82,28%
P. Sujatha and K. Mahalakshmi [47]	2020	- RF	83,51%
		- LR	80,21%
		- DT	79,90%
		- KNN	72,52%
Singh and R. Kumar [48]	2020	- KNN	87,01%
Sateesh.R, Thomas.A [49]	2020	- CNN	100%
R. Sateesh, S. Sameen [50]	2021	- E KNN	98%
		- SVM	89%
		- KNN	91,92%
		- DT	87,91%
Notre approche	2022	-RF / GA-RF	89.2% / 89.28%
		-DT / GA-DT	83.9% / 78.57%
		-LR / GA-LR	87.5% / 87.5%
		-SVM / GA-SVM	87.5% / 87.5%
		-KNN / GA-KNN	92.8% / **96.42**%

TABLE 3.1 – Comparaison des travaux étudiés.

3.3 Approche et solution proposées

Nous illustrons notre approche dans la figure 3.1, elle comprend 7 étapes, tel que la première étape est désignée comme la collecte des données, la deuxième étape est l'ana-

lyse exploratoire des données qui sert à comprendre au maximum les données dont on dispose pour définir une stratégie de modélisation, la troisième étape est le pré traitement des données, dont on fait le filtrage des valeurs manquantes, la suppression des valeurs redondantes et aberrantes, et la normalisation des données .

Après le pré traitement des données, nous avons choisi de faire la sélection des attributs afin de trouver le meilleur vecteur d'attributs pertinents de l'ensemble de données en utilisant l'algorithme génétique. Ensuite nous divisons l'ensemble de données avec la méthode train-test-split à un ensemble d'entraînement et un autre pour le test afin d'appliquer les modèles d'apprentissage automatiques supervisés choisis (KNN, RF, LR, SVM et DT).

Enfin, nous évaluons nos modèles en utilisant la matrice de confusion et les métriques d'évaluation de performances (Accuracy, F1-score, Précision, Rappel) pour choisir le meilleur modèle selon son taux de prédiction.

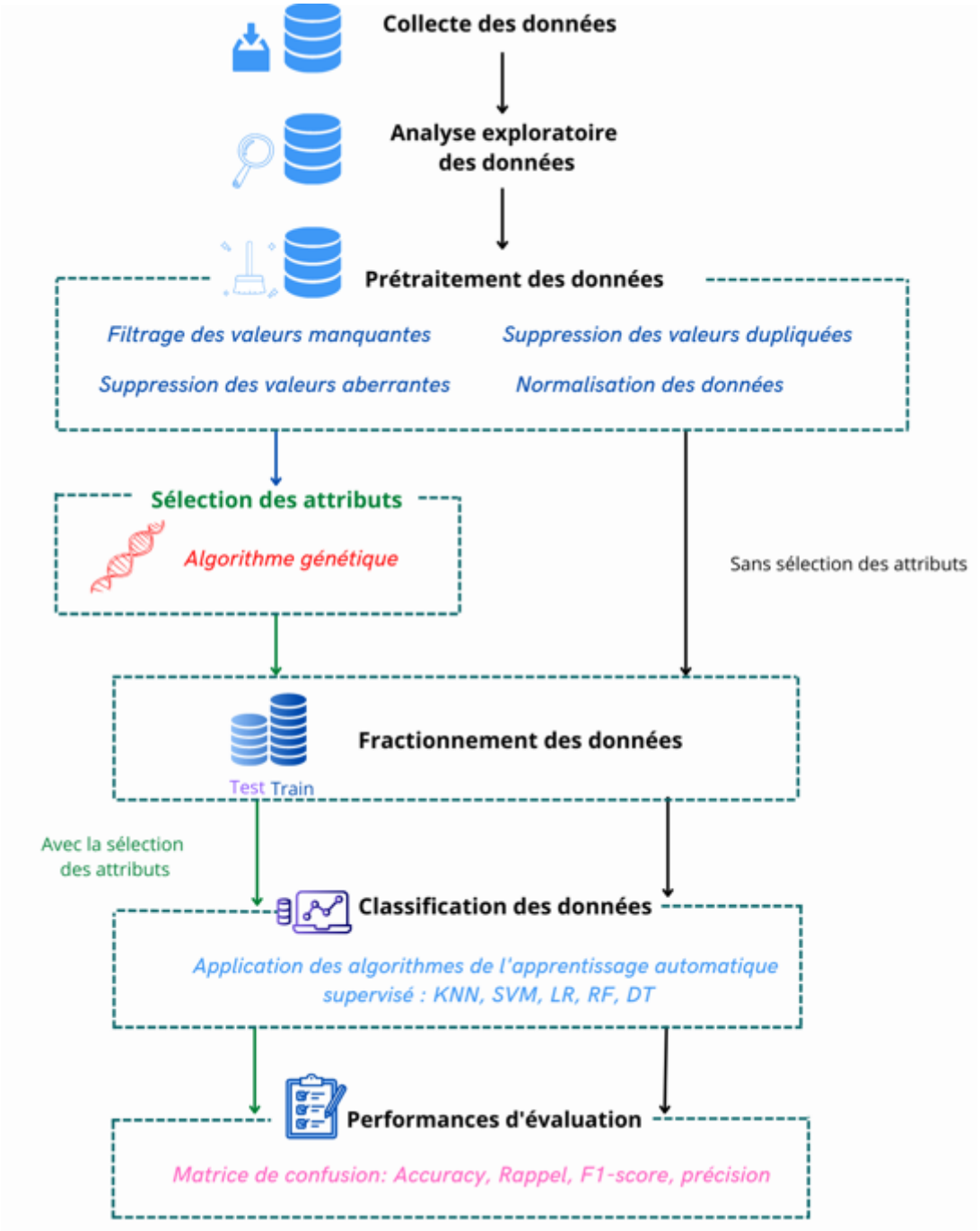

FIGURE 3.1 – Architecture de système proposé.

3.4 Collecte des données

Cette étape consiste à rassembler les données nécessaires pour l'apprentissage automatique d'une façon consolidée, afin qu'elles soient toutes contenues dans un seul tableau. Dans notre étude nous utilisons l'ensemble de données Cleveland [50] cet ensemble contient 76 attributs, mais toutes les expériences publiées se réfèrent à l'utilisation d'un

sous-ensemble de 14 d'entre eux, et 303 de patients. la base de données de Cleveland est la seule à avoir été utilisée par les chercheurs en apprentissage automatique à ce jour, Le champs "target" fait référence à la présence ou l'absence d'une maladie cardiaque chez le patient.

Le tableau 3.2 montre les caractéristiques cliniques et leur description de la base de données de Cleveland :

	Attributs	**Description**	**Type**
1	Age	Age du patient (29 à 77)	numérique
2	Sexe	Le sexe de la personne (0 :féminin, 1 :masculin)	catégorique
3	Cp	Le type de douleur thoracique est classé en quatre types 1) une douleur causée par une angine typique, 2) causée par une angine atypique, 4) une douleur non bouchante 5) asymptomatique	catégorique
4	Trestbps	Tension artérielle au repos (en mm Hg à l'admission à l'hôpital)	numérique
5	Chol	Cholestérol sérique en mg / dL(126 à 564)	numérique
6	Fbs	Glycémie à jeun >120 mg / dL(1 : vrai,0 : faux)	catégorique
7	Restecg	Résultat électrocardiographique au repos (0 à 2)	catégorique
8	Thalache	La fréquence cardiaque maximale de la personne atteinte (71 à 202)	numérique
9	Exang	Angine induite par l'exercice. est une condition où pas assez de sang est fourni aux parois du coeur pour pomper le sang. Il est causé par exercice ou tout stress physique ou mental (oui : 1,non : 0)	catégorique
10	Oldpeak	ST dépression induite par l'exercice par rapport au repos ('ST' se rapporte aux positions sur le graphique ECG.)	numérique
11	Slope	la pente du segment ST peak exercice (0 à 1)	catégorique
12	Ca	Le nombre de grands vaisseaux (0 à 3)	numérique
13	Thal	Un trouble sanguin appelé thalassémie	catégorique
14	Target	présence ou absence de maladie cardiaque (1 ou 0)	catégorique

TABLE 3.2 – Description des variables d'ensemble de données.

3.5 Analyse exploratoire des données

L'analyse exploratoire des données (AED) est un processus ouvert dans le cadre du quel nous calculons des statistiques et établissons des chiffres pour trouver des tendances, des anomalies, des modèles ou des relations entre les données. Afin de comprendre au maximum les données dont on dispose pour définir une stratégie de modélisation.
Elle dispose deux parties, analyse de la forme et analyse de fond.

3.5.1 Analyse de la forme

La figure 3.2 nous permet d'extraire les informations de base sur l'ensemble de données telles que :

- **La taille** : Il se compose de 303 lignes et 14 colonnes y compris le target.
- **Les colonnes** : Les colonnes de l'ensemble de données sont : ['age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach','exang', 'oldpeak', 'slope', 'ca', 'thal', 'target'].
- **Le type** : L'ensemble de données et de type : Object.
- **Le détails statistique** : Il nous fournit des informations statistiques au format numérique. nous pouvons en déduire que dans la colonne AGE, l'âge minimum est de 29 ans et le maximum est de 77 ans, l'âge moyen est de 54 ans. Les détails des quartiles sont donnés sous forme de 25%, 50% et 75%. cela nous aide pour avoir plus d'informations pour la prochaine étape Analyse de fond (voir section 3.5.2).

```
df.shape
```

```
(303, 14)
```

```
df.columns
```

```
Index(['age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach',
       'exang', 'oldpeak', 'slope', 'ca', 'thal', 'target'],
      dtype='object')
```

```
df.describe()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	2.313531	0.544554
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	0.612277	0.498835
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

FIGURE 3.2 – Description de l'ensemble de données.

Les types d'attributs sont :

— **Attributs qualitatives (binaires ou catégoriques)** : sont des attributs à deux catégories ou plus et chaque valeur de cette caractéristique peut être classés par eux) : [sexe, fbs, exang, cible, cp, restecg ,slope, ca, thal]

— **Attributs quantitatives (continus ou numériques)** : sont des attributs prenant des valeurs entre deux points quelconques ou entre le minimum et maximum de valeurs dans la colonne de caractéristique) : [âge, trestbps, chol, thalach, oldpeak]

3.5.2 Analyse de fond

3.5.2.1 Analyse uni-variée

L'analyse uni-variée a pour but de décrire et de mesurer la répartition des valeurs qui peut prendre une variable dans l'ensemble de données.

Visualisation de la classe cible

La visualisation de la classe cible est importante pour voir l'équilibrage des données qui est essentiel pour avoir un résultat précis, La Figure 3.3 représente les graphiques de visualisation de la classes cible. elle nous montre que le nombre de personnes sans maladie cardiaque est de 138 avec une proportion de $44,14\%$ et le nombre de personnes atteintes de maladie cardiaque est de 165 avec une proportion de $55,86\%$. Nous notons que le pourcentage de personnes atteintes de maladies cardiaques et de personnes sans maladie est presque proche c'est-à-dir un dataset presque équilibré.

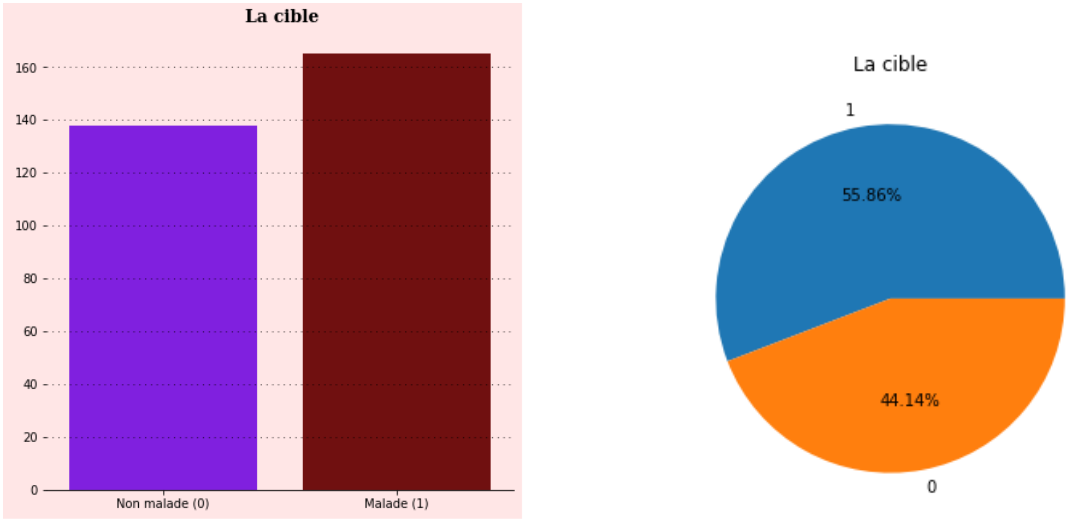

FIGURE 3.3 – Visualisation de la classe cible.

Étude des variables qualitatives (catégoriels)

La figure 3.4 représente les attributs catégoriels de notre ensemble de données.

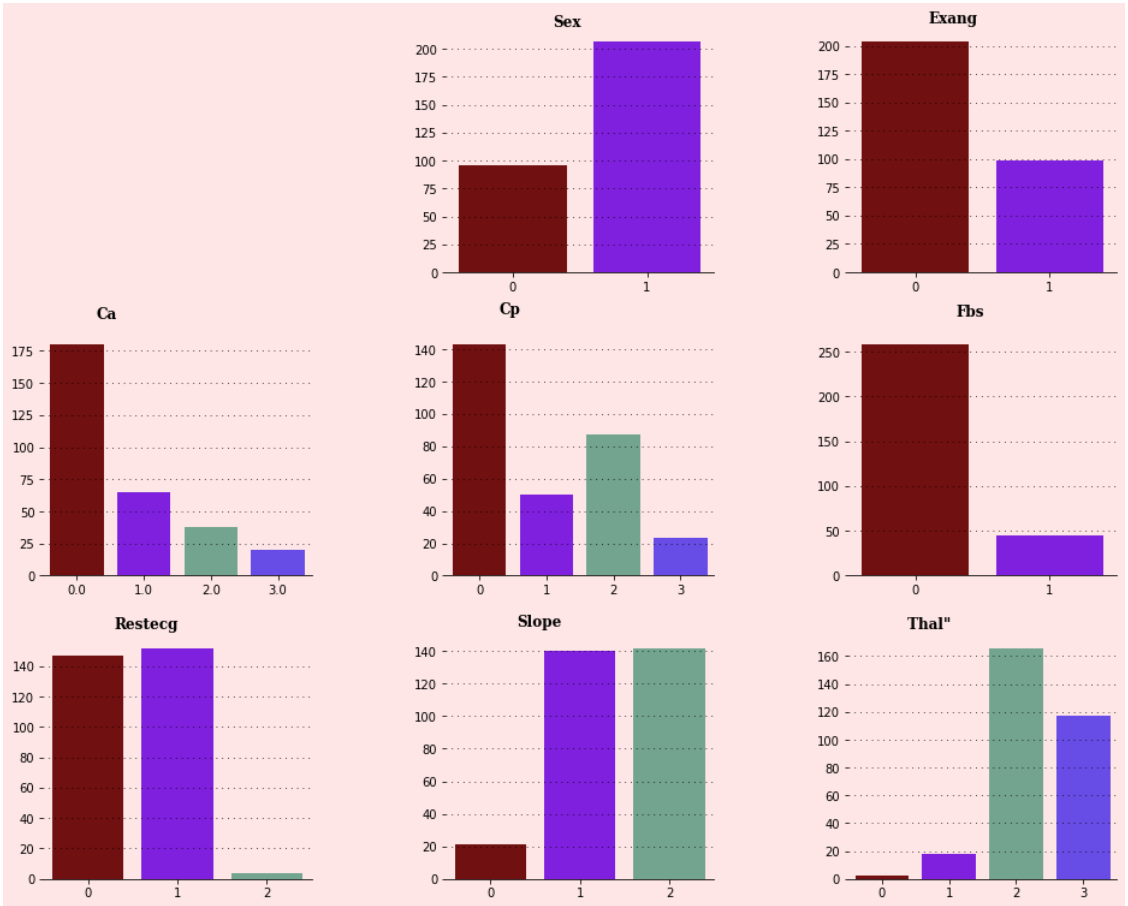

FIGURE 3.4 – Visualisation des attributs qualitatives "catégoriels".

Les résultats montrent les informations suivantes :

— **Sex :** Dans l'ensemble de données, les hommes sont majoritairement représentés que les femmes.

— **Cp :** Il y a plus de 140 personnes qui ont des douleurs thoraciques asymptomatiques, 50 personnes qui ont des douleurs non angineuses, 90 personnes qui ont une angine atypique et 25 personnes qui ont une angine typique.

— **Restecg :** Presque la moitié des individus ont une activité électrique du coeur assez normale (145 personnes).

— **Exang :** plus de 200 personnes del'ensemble de données ont une angine induite par l'exercice.

— **Slope** La moitié des individus ont une pente plate du segment ST. 140 individus ayant une pente ascendante. 22 personnes restantes ont une pente descendante.

— **Fbs :** Une grande partie des individus plus de 250 personnes ont une glycémie à jeun supérieure à 120 mg/dl.

— **Thal :** Une majorité d'individus (environ 160) a le trouble sanguin thalassémie de type 2 as-symptomatique.

Étude de variables quantitatives

La figure 3.5 représente les attributs quantitatives de notre ensemble de données.

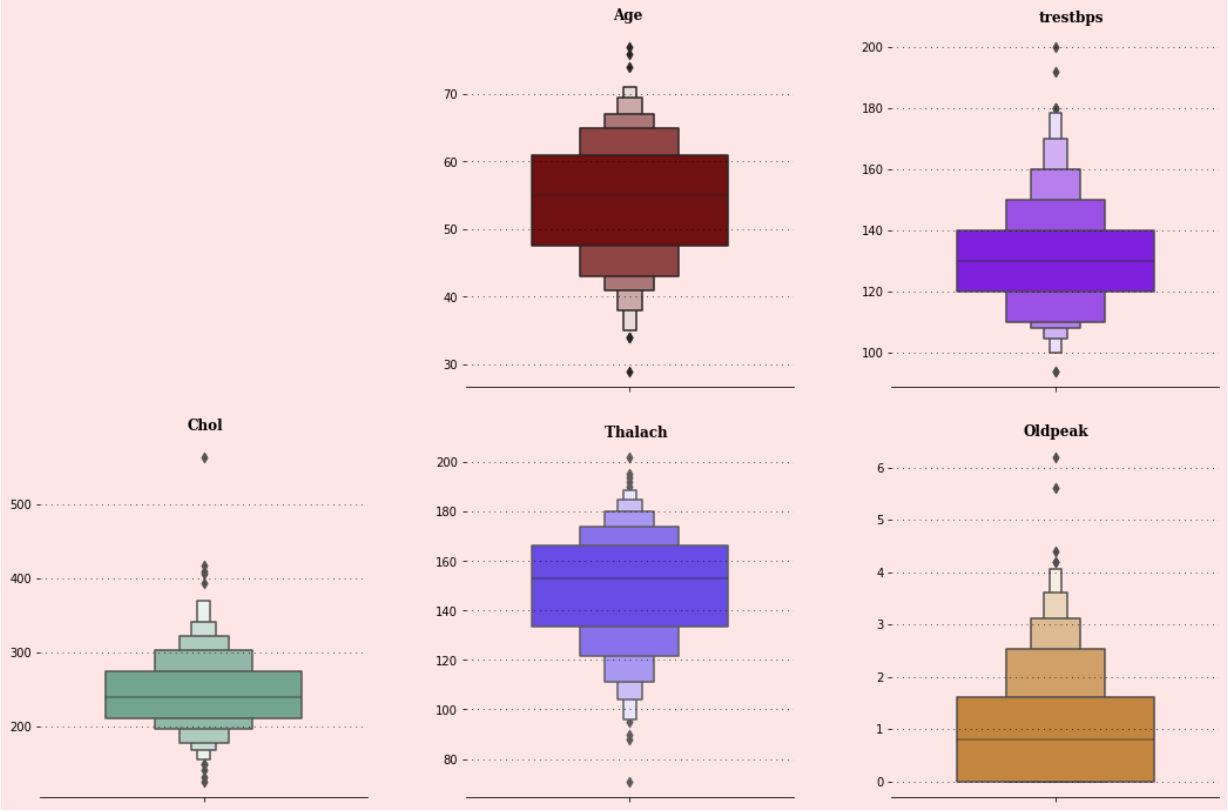

FIGURE 3.5 – Visualisation des attributs quantitatives.

Les résultats de la figure 3.5 montrent les informations suivantes :

— **Age :** Nous somme en présence d'une population âge moyen. Âge compris entre 45 et 65 ans, Le plus jeune à 28 ans tandis que le plus âgé à 77 ans, L'âge moyen est de 54 ans.

— **Thalach :** Le rythme cardiaque le plus bas est de 60 battements/minute et le plus élevé est 202 battements/minutes ; la majorité des personnes ont un rythme cardiaque de 150 battements/minutes ;

— **Oldpeak :** La tension artérielle au repos le plus bas est de 0 unité de mesure et le plus élevé est 200 unité de mesure ; la majorité des personnes ont une pression artérielle au repos de 132,39 ;

— **Dépression ST :** Distribution étalée en bas, 50% des individus ont une dépression inférieur à 0.9 (Moyenne > mediane).

— **Cholesterol :** Distribution étalée à droite, 50% des individus ont une taux de cholestérol supérieur à 199 (Moyenne < mediane).

— À travers les différentes boîtes BoxenPlots, on remarque que les variables ['thalach', trestbps', 'chol', 'oldpeak'] contiennent des valeurs aberrantes. nous allons les gérer dans la section 3.6.3.

3.5.2.2 Analyse bivariée

L'analyse bivariée est l'une des formes les plus simples d'analyse quantitative et qualitative. Elle implique l'analyse de deux variables, afin de déterminer la relation empirique entre elles. L'analyse bivariée peut être utile pour tester des hypothèses d'association simples [51].

Visualisation de la relation variables qualitatives /classe cible

La Figure 3.6 représente la répartition des caractéristiques catégorielles selon la classe cible. On remarque que les femmes, les personnes ayant une douleur thoracique atypique (Cp_2), celles qui n'ont pas une angine après un exercice et celles dont la pente du segment ST est plate, sont beaucoup plus atteintes par les maladies cardiaques.
Nous avons testé ces hypothèses avec le test de *khi-carré* (Annexe A, section A.1.1).

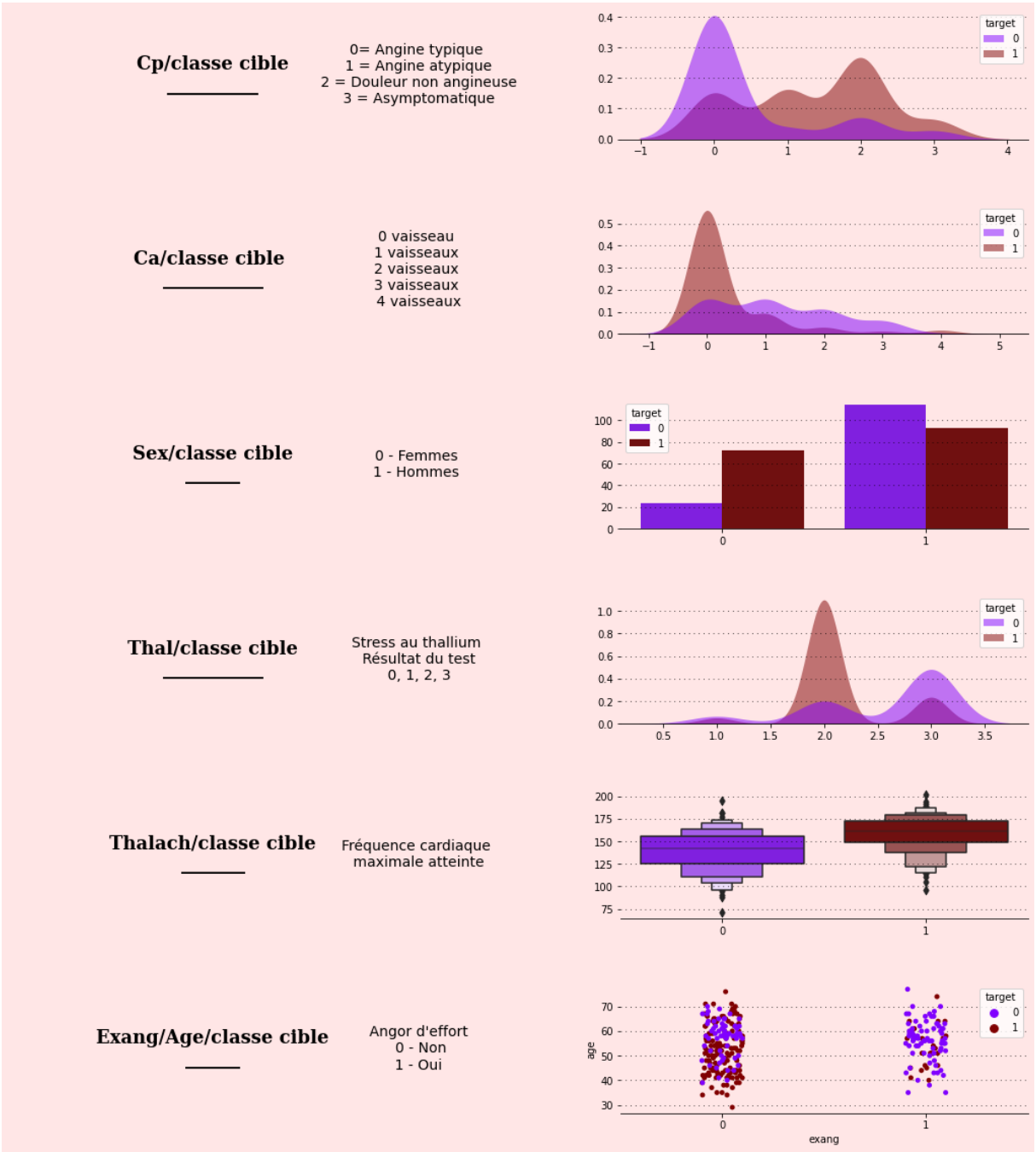

FIGURE 3.6 – Répartition des caractéristiques catégorielles selon la classe cible.

La tableau 3.3 montre les résultats de test khi-carré sur les variables qualitatives de l'ensemble de données, les résultats montrent que $p - value[fbs] > 5\%$ ce qui montre qu'elle est indépendante de l'ensemble de données donc n'est pas si importante pour le diagnostique des maladies cardiaques.

Attributs	P-value
'ca'	0.0
'cp'	0.0
'exang'	0.0
'fbs'	**0.7444**
'restecg'	0.0067
'sex'	0.0
'slope'	0.0
'thal'	0.0

TABLE 3.3 – Résultats de test statistique Khi-carré.

Visualisation de la relation variables quantitatives /classe cible

La Figure 3.7 nous constate que les personnes susceptibles d'avoir une maladie cardiaque sont les personnes moyennement âgées ou celles qui ont une fréquence cardiaque maximale basse ou encore celles qui ont une dépression élevée.
La plupart des autres variables semblent de ne pas influencer le fait qu'une personne soit malade ou pas.

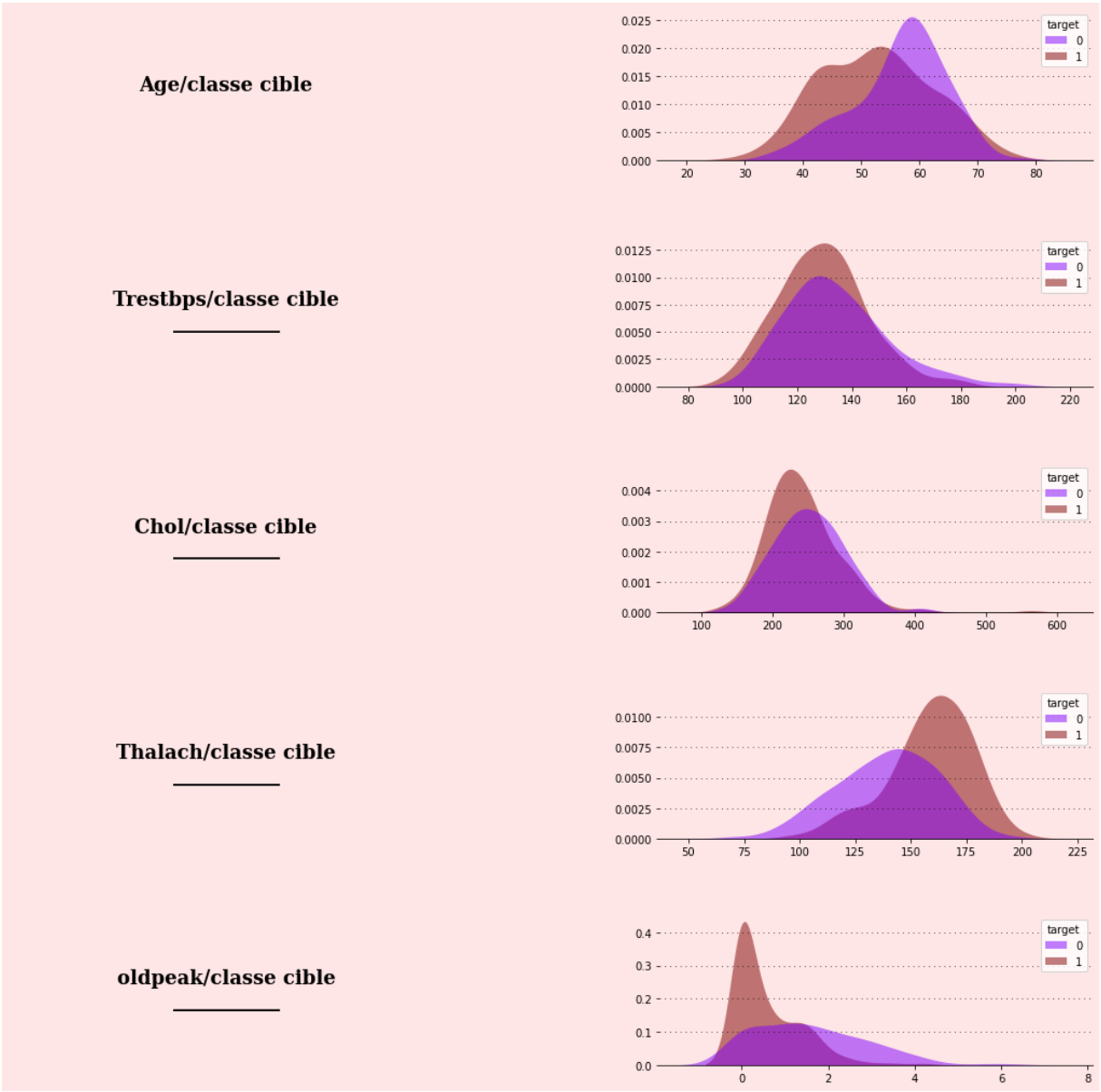

FIGURE 3.7 – Répartition des caractéristiques quantitatives selon la classe cible.

La question est de savoir si une combinaison de plusieurs variables peuvent influencer le fait qu'une personne soit malade ou pas, nous avons tester ces hypothèses avec une analyse de la variance avec le test *ANOVA* (Annexe A, section A.1.2). Le tableau 3.4 montre que les variables [chol, thalach, oldpeak] ne sont pas corrélées avec la cible.

Attributs	P-value
'Age'	0.1192
'trestbps'	0.4998
'chol'	**0.0394**
'thalach'	**0.0428**
'oldpeak'	**0.0**

TABLE 3.4 – Résultats de test statistique Anova.

3.5.2.3 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 un 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 p [52].

— Plus p est proche de zéro, plus la relation linéaire est faible.

— Les valeurs positives de p indiquent une corrélation positive lorsque les valeurs des deux variables tendent à augmenter ensemble.

— Les valeurs négatives de p indiquent une corrélation négative lorsque les valeurs d'une variable tend à augmenter et que les valeurs de l'autre variable diminuent.

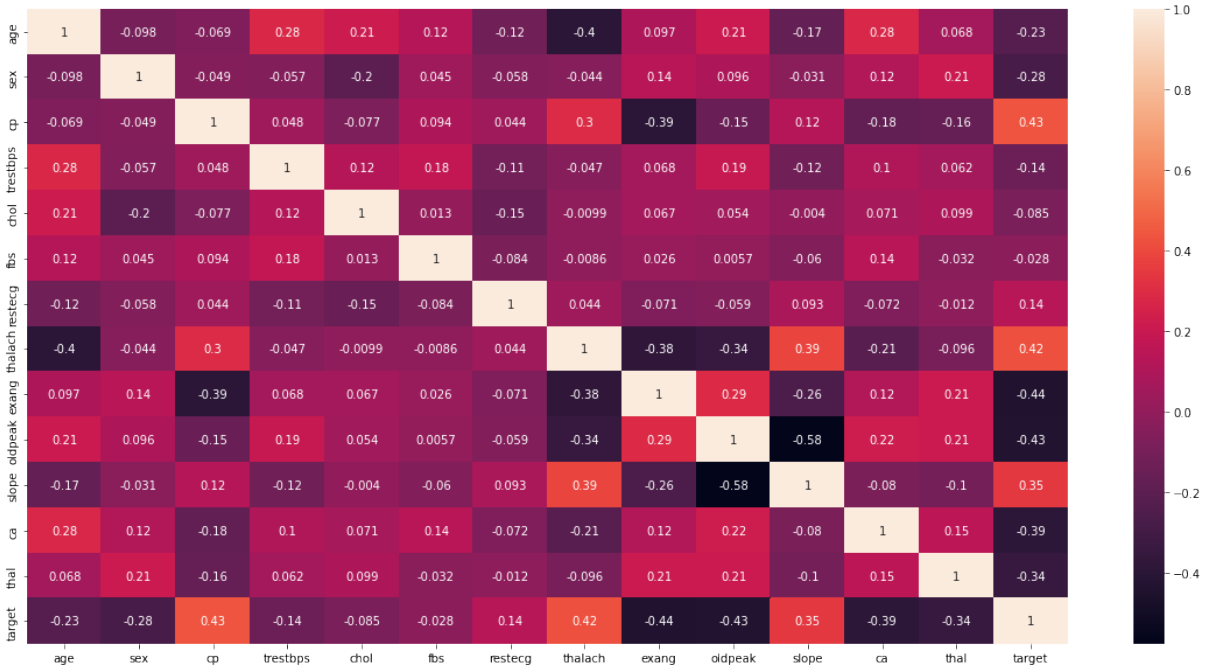

FIGURE 3.8 – Matrice de corrélation.

À travers la matrice de corrélation illustré dans la figure 3.8, on constate qu'il y a une très faible corrélation entre les variables quantitatives [âge, trestbps, chol, thalach, oldpeak] lorsqu'elles sont croisées deux à deux. Les valeurs des coefficients de corrélation sont dans l'intervalle$[-0.38, 0.26]$. Nous pouvons dire que les attributs de notre ensemble de données ne sont pas corrélés.

3.6 Pré-traitement des données

L'objectif de cette étape est de mettre les données dans un format propice aux systèmes d'apprentissage automatique afin d'améliorer la performance de ses modèles et peut se faire comme suit :

3.6.1 Filtrage des valeurs manquantes

Filtrage des valeurs manquantes (NaN) consiste à remplacer les (NaN) par d'autres valeurs (la moyenne de la série, le médian de la série, la moyenne des voisins,... etc) ou les supprimer et s'assurer de ne pas avoir de doublons.

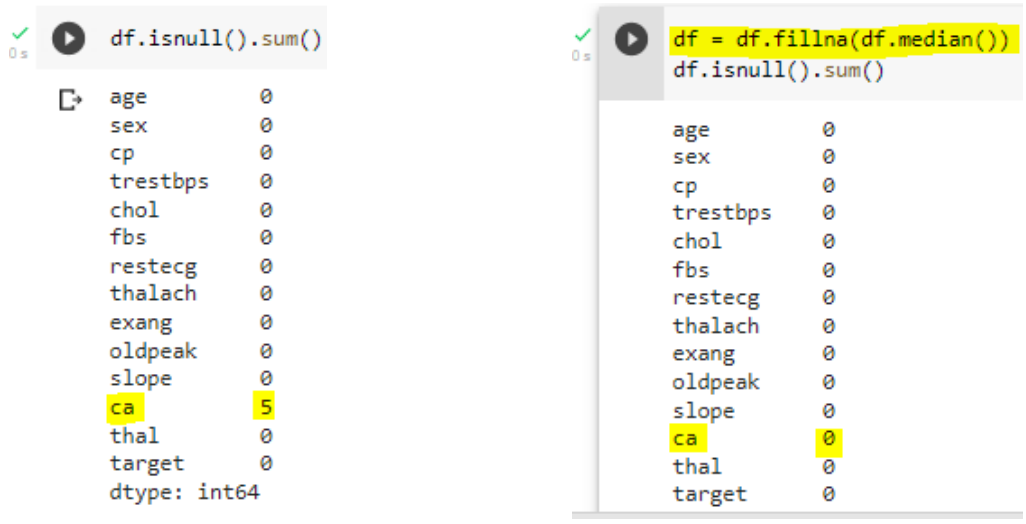

FIGURE 3.9 – Les valeurs manquantes (NaN).

La figure 3.9 montre que l'ensemble de données a 5 valeurs nulles dans la variable ['ca'], nous pouvons visualiser ces valeurs grâce à la bibliothèque missingo (voir figure 3.10), donc nous remplaçons les NaNs par la mediane.

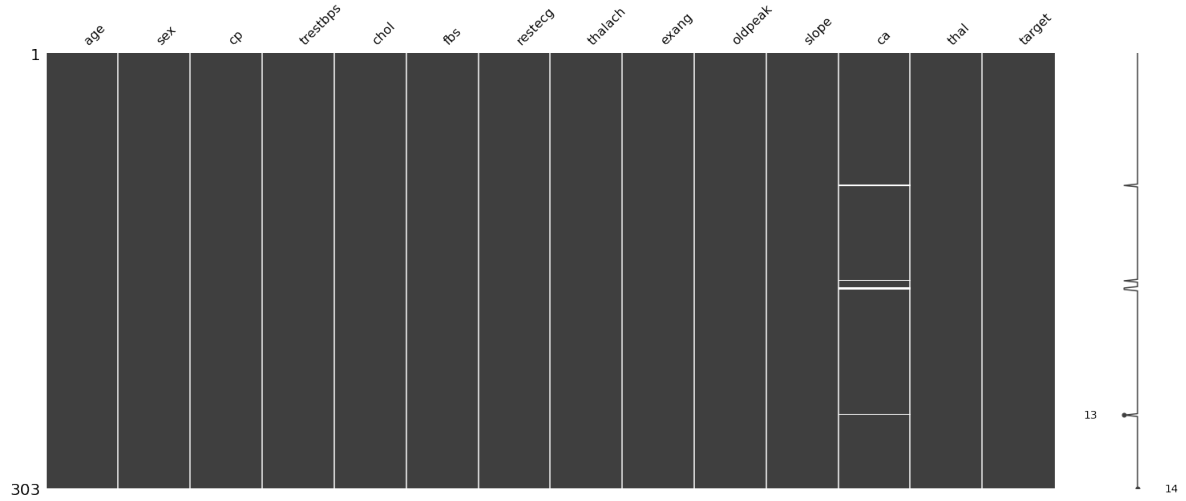

FIGURE 3.10 – Visualisation des valeurs manquantes.

3.6.1.1 Suppression des valeurs uniques

Nous avons utiliser la fonction de DataFrame Panda *nunique()* pour trouver le nombre de valeurs uniques sur l'axe d'index dans notre ensemble de données, nous avons obtenue les résultats suivants :

— La variable ['ca'] varie de 0 à 3 , cependant, son *df.nunique()* = 5 répertoriée de 0 à 4 . donc nous trouvons les valeurs de la catégorie '4' et changeons-les en *NaN* (voir figure3.11).

```
[39] df['ca'].unique()

     array([0, 2, 1, 3, 4])

[40] df.ca.value_counts()

     0    175
     1     65
     2     38
     3     20
     4      5
     Name: ca, dtype: int64

[41] df[df['ca']==4]
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
92	52	1	2	138	223	0	1	169	0	0.0	2	4	2	1
158	58	1	1	125	220	0	1	144	0	0.4	1	4	3	1
163	38	1	2	138	175	0	1	173	0	0.0	2	4	2	1
164	38	1	2	138	175	0	1	173	0	0.0	2	4	2	1
251	43	1	0	132	247	1	0	143	1	0.1	1	4	3	0

```
df.loc[df['ca'] == 4, 'ca'] = np.NaN
df['ca'].unique()

array([ 0.,  2.,  1.,  3., nan])
```

FIGURE 3.11 – Correction des valeurs uniques de ['ca'].

— La variable ['thal'] varie de 1 à 3 , cependant, son *df.nunique()* = *4* répertorié 0–3. Il y a deux valeurs de '0' . Alors changeons-les en *NaN* Voir figure 3.12.

```
df['thal'].unique()
```

```
array([1, 2, 3, 0])
```

```
df.thal.value_counts()
```

```
2    166
3    117
1     18
0      2
Name: thal, dtype: int64
```

```
df[df['thal']==0]
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
48	53	0	2	128	216	0	0	115	0	0.0	2	1.0	0	1
281	52	1	0	128	204	1	1	156	1	1.0	1	1.0	0	0

```
df.loc[df['thal'] == 0, 'thal'] = np.NaN
df['thal'].unique()
```

```
array([ 1.,  2.,  3., nan])
```

FIGURE 3.12 – Correction des valeurs uniques de ['thal'].

3.6.2 Suppression des valeurs dupliquées

Nous avons trouvé une seule ligne dupliquée dans notre ensemble de données, Les résultats sont montrés dans la figure 3.13, la taille de l'ensemble de données est devenue (302,14)

```
duplicated = df.duplicated().sum()
if duplicated:
  print("Nbr de lignes dupliquées dans l'ensemble de données est {}".format(duplicated))
else:
  print("l'ensemble de données ne contient pas des lignes dupliquées")
```

```
Nbr de lignes dupliquées dans l'ensemble de données est 1
```

```
du=df.drop_duplicates(keep='first')
du.shape
```

```
(302, 14)
```

FIGURE 3.13 – Suppression des valeurs dupliquées.

3.6.3 Suppression des valeurs aberrantes (Outliers)

Plusieurs algorithmes d'apprentissage automatique sont sensibles aux données d'entraînement ainsi qu'à leurs distributions. Avoir des valeurs abberantes (outliers en anglais) dans l'ensemble d'entraînement d'un algorithme peut rendre la phase d'entraînement plus longue. Sans mentionner que l'apprentissage sera biaisé. Par conséquent, le modèle prédictif produit ne sera pas performant, ou du moins, loin d'être optimal [53].
Dans la figure 3.5 à l'aide des boxplot nous observons l'existence des valeurs aberrantes dans notre ensemble de données dans les colonnes suivantes : ['thalach', trestbps', 'chol', 'oldpeak'], On supprime 24 valeurs aberrantes, la taille de l'ensemble de données est devenue (278,14).

```
Outliers: [71, 71, 71]
Les valeurs aberrantes dans [age] sont [71, 71, 71]
Outliers: [180, 180, 180]
Les valeurs aberrantes dans [trestbps] sont [180, 180, 180]
Outliers: [141, 149, 149, 354, 360, 394]
Les valeurs aberrantes dans [chol] sont [141, 149, 149, 354, 360, 394]
Outliers: [4.0, 4.2, 4.2]
Les valeurs aberrantes dans [oldpeak] sont [4.0, 4.2, 4.2]
Outliers: [96, 96, 97, 187, 188, 190]
Les valeurs aberrantes dans [thalach] sont [96, 96, 97, 187, 188, 190]
```

```
[98] du.drop((du[du.chol > 400].index) | (du[du.chol < 141].index), inplace=True)
     du.drop((du[du.thalach > 191].index) | (du[du.thalach < 96].index), inplace=True)
     du.drop((du[du.trestbps > 191].index) | (du[du.trestbps < 95].index), inplace=True)
     du.drop((du[du.age > 73].index) | (du[du.age < 35].index), inplace=True)
     du.drop(du[du.oldpeak > 4.3].index, inplace=True)
     du.shape

     (278, 14)
```

FIGURE 3.14 – Suppression des valeurs aberrantes.

3.6.4 Normalisation des données

3.6.4.1 Transformation des valeurs catégoriques à des valeurs numériques

Avant d'aller plus loin, nous devons traiter des variables catégoriques car un modèle d'apprentissage automatique ne peut malheureusement pas traiter les variables catégoriques, avec d'autres numériques, il existe plusieurs méthodes de mener ce processus

comme :

- LabelEncoder : attribuer chaque catégorie unique dans une variable catégorielle avec un entier.
- One-hot encoding : créer une nouvelle colonne pour chaque catégorie unique dans une variable catégorielle.
- Dummies-pandas methode : elle consiste à créer une colonne séparée pour chacune des valeurs uniques des colonnes de catégories. Comme la valeur de chaque colonne est binaire (0/1), nous ne pouvons donc avoir qu'une seule valeur 1 dans les colonnes nouvellement générées.

Nous avons choisit la troisième méthode, Les résultats de la figure : 3.15 sont faite par le code python suivant :

```
du = pd.get_dummies(du, columns = ['ca','cp','exang', 'fbs', 'restecg'
,'sex','slope', 'thal'])
```

	age	trestbps	chol	thalach	oldpeak	target	ca_0	ca_1	ca_2	ca_3	...	restecg_2	sex_0	sex_1	slope_0	slope_1	slope_2	thal_0	thal_1	thal_2	thal_3
0	63	145	233	150	2.3	1	1	0	0	0	...	0	0	1	1	0	0	0	1	0	0
1	37	130	250	187	3.5	1	1	0	0	0	...	0	0	1	1	0	0	0	0	1	0
2	41	130	204	172	1.4	1	1	0	0	0	...	0	1	0	0	0	1	0	0	1	0
3	56	120	236	178	0.8	1	1	0	0	0	...	0	0	1	0	0	1	0	0	1	0
4	57	120	354	163	0.6	1	1	0	0	0	...	0	1	0	0	0	1	0	0	1	0

5 rows × 31 columns

FIGURE 3.15 – Mise en forme et transformation des colonnes catégoriques à des colonnes numériques.

3.6.4.2 Mise à l'échelle des données

Étant donné que la plage de valeurs des données brutes varie considérablement, dans certains algorithmes d'apprentissage automatique, les fonctions objectives ne fonctionneront pas correctement sans normalisation. Par exemple, de nombreux classificateurs calculent la distance entre deux points par la distance euclidienne (dans l'algorithme KNN). Si l'une des caractéristiques a une large plage de valeurs, la distance sera régie par cette caractéristique particulière. Par conséquent, la plage de toutes les caractéristiques doit être normalisée afin que chaque caractéristique contribue approximativement proportionnellement à la distance finale.

Il existe plusieurs méthodes pour faire la normalisation des données : Remise à l'échelle (normalisation min-max), Mean normalization, Standardization (Z-score Normalization). Nous avons utilisé le $Standar\_Scaller$ de la bibliothèque $scikit-learn$ avec le code python suivant :

```
change_scale = ['age', 'chol', 'oldpeak', 'thalach', 'trestbps']
du[change\_scale] = StandardScaler().fit_transform(du[change_scale])
```

La figure 3.16 montre les résultats de la normalisation sur notre ensemble de données :

	age	trestbps	chol	thalach	oldpeak	target	ca_0	ca_1	ca_2	ca_3	...	restecg_2	sex_0	sex_1	slope_0	slope_1	slope_2	thal_0	thal_1	thal_2	thal_3
0	1.007439	0.823153	-0.268841	0.000169	1.216366	1	1	0	0	0	...	0	0	1	1	0	0	0	1	0	0
1	-2.000313	-0.078015	0.111106	1.742383	2.336375	1	1	0	0	0	...	0	0	1	1	0	0	0	0	1	0
2	-1.537582	-0.078015	-0.916986	1.036080	0.376358	1	1	0	0	0	...	0	1	0	0	0	1	0	0	1	0
3	0.197659	-0.678793	-0.201792	1.318601	-0.183647	1	1	0	0	0	...	0	0	1	0	0	1	0	0	1	0
4	0.313342	-0.678793	2.435489	0.612298	-0.370315	1	1	0	0	0	...	0	1	0	0	0	1	0	0	1	0

5 rows × 31 columns

FIGURE 3.16 – Normalisation des données.

3.7 Sélection des attributs avec l'algorithme génétique

La sélection des attributs est l'étape de base de notre étude. Nous avons utilisé la recherche génétique comme mesure de qualité pour éliminer les attributs redondants et non pertinents, et pour classer les attributs qui contribuent le plus à la classification. Les derniers attributs classés sont supprimés et L'algorithme de classification est construit sur la base des attributs évalués.

3.7.1 Algorithme génétique

Le principe de fonctionnement de l'algorithme génétique est déjà expliqué dans la section 2.5.2.1 dans chapitre 2, Les étapes de cet algorithme sont illustrées dans le graphe illustré de la figure 3.17 :

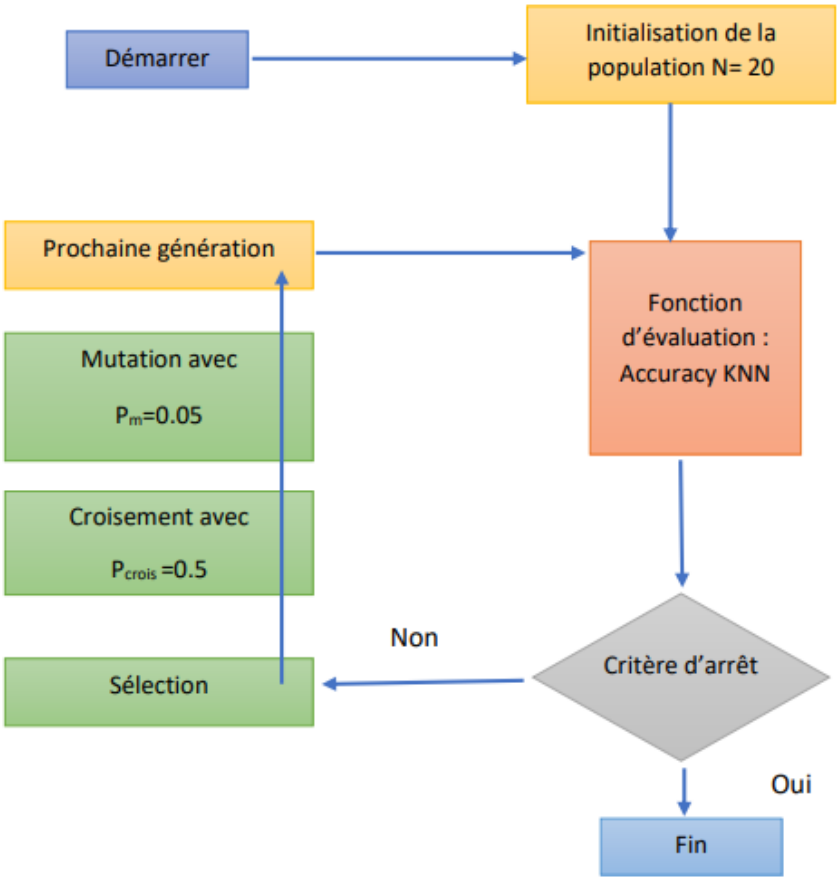

FIGURE 3.17 – Étapes de l'algorithme génétique.

Nous présenterons dans ce qui suit l'application de l'algorithme génétique sur un exemple de notre étude, la partie expliquée dans cet exemple est juste la première production car l'algorithme travail en boucle.

3.7.1.1 Étape 1 : Codage

Nous avons choisit d'utiliser le codage binaire pour le codage de nos chromosomes (13 bits = 13 attributs), son principe est de coder la solution selon une chaîne de bits (qui peuvent prendre les valeurs 0 = attribut pertinent ou 1 = attribut non pertinent) [54].

3.7.1.2 Étape 2 : Initialisation de la population

Le but de cette étape est de produire une population d'individus non homogène, le choix de cette population est aléatoire .

Donc nous fixons la taille de la population à $N = 20$ (la taille de la population est choisit par expérimentation). Nous tirons donc de façon aléatoire 4 chromosomes sachant qu'un chromosome est composé de 13 bits et que chaque bit dispose d'une probabilité $1/2$ d'avoir une valeur 0 ou 1 [54].

Num	Chromosomes	Codage binaire
1	[Age, Sex, trestbps, chol, Fbs, Restecg, exang, oldpeak, thal]	1 1 0 1 1 0 1 0 1 1 0 0 1
2	[Age, Cp, trestbps, chol, thalach, oldpeak, thal]	1 0 1 1 1 0 0 1 0 1 0 0 1
3	[Sex, cp, trestbps, restecg, thalach, oldpeak, ca, thal]	0 1 1 1 0 0 1 1 0 0 0 1 0
4	[Age, sex, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal]	1 1 0 0 1 1 1 1 1 1 1 1 1

TABLE 3.5 – Exemple de 4 chromosomes.

3.7.1.3 Étape 3 : Calcule de la fonction d'évaluation (fitness)

La fonction *fitness* est la partie la plus importante de l'AG parce que une bonne *fitness* peut améliorer le résultat de l'AG. Nous utilisons comme fonction *fitness* le taux d'accuracy du *KNN*.

Après l'application de la fonction *fitness* sur l'individu choisit (les 4 chromosomes choisis aléatoirement) nous remarquons que le maximum est : 0.89 est atteint par le quatrième chromosome (voir tableau 3.6).

N°	Chromosomes	Codage binaire	*Fitness*
1	[Age, Sex, trestbps, chol, Fbs, Restecg, exang, oldpeak, thal]	1 1 0 1 1 0 1 0 1 1 0 0 1	0.801
2	[Age, Cp, trestbps, chol, thalach, oldpeak, thal]	1 0 1 1 1 0 0 1 0 1 0 0 1	0.732
3	[Sex, cp, trestbps, restecg, thalach, oldpeak, ca, thal]	0 1 1 1 0 0 1 1 0 0 0 1 0	0.751
4	[Age, sex, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal]	1 1 0 0 1 1 1 1 1 1 1 1 1	**0.890**

TABLE 3.6 – Évaluation de l'individu avec la fonction $fitness$.

3.7.1.4 Étape 4 : Sélection

La sélection consiste à choisir les individus à partir des quels on va créer la génération suivante. La sélection des individus s'effectue le plus souvent sur la base de leur fonction d'évaluation [54].

Plusieurs opérateurs de sélection existent :

- Les méthodes par probabilité (à la roulette Goldberg et David [55])
- Par rang de classement dans la population (Davis [55])
- Par tournoi (Miller et Goldberg [55])

Dans notre exemple d'étude, nous avons choisi le processus de sélection à la roulette, cette technique est inspirée des roues de loterie. chaque individus de la population est associé un secteur d'une roue. L'angle du secteur étant proportionnel à la qualité de l'individu qu'il représente. à chaque fois nous tournons la roule on tire un chromosome. pour chaque chromosome x avec une valeur de fitness correspondante f(x), nous calculons la probabilité $p_s(x)$de sélection correspondante comme :

$$p_s(x) = \frac{f(x)}{\sum_{i=1}^{n} f_i(x)} \tag{3.1}$$

Nous obtenons les résultats dans le tableau 3.7 et la roue de loterie dans la figure 3.18.

N°	Chromosomes	f(x)	$p_s(x)$
1	1 1 0 1 1 0 1 0 1 1 0 0 1	0.801	0.252
2	1 0 1 1 1 0 0 1 0 1 0 0 1	0.732	0.230
3	0 1 1 1 0 0 1 1 0 0 0 1 0	0.751	0.236
4	1 1 0 0 1 1 1 1 1 1 1 1 1	0.890	0.280

TABLE 3.7 – Propabilité de sélection.

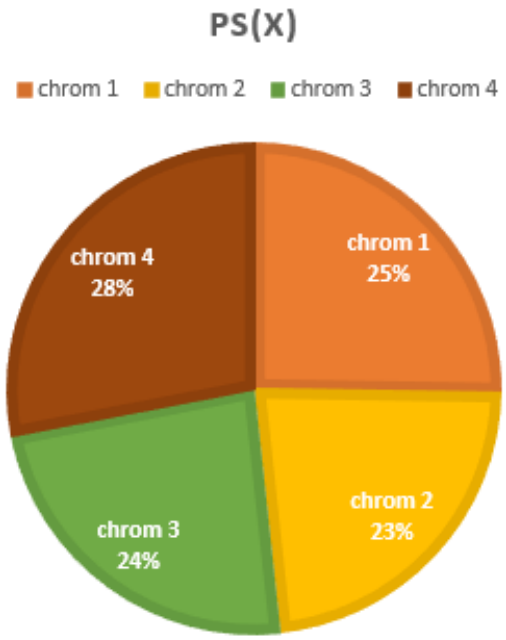

FIGURE 3.18 – Résultats de la sélection à la roulette.

3.7.1.5 Étape 5 : Croisement

Nous avons appliqué ce processus à chaque paire de chromosomes sélectionnés avec la probabilité $P_{crois} = 0.5$. c'est à dire 50% d'enfants seront une copie exacte des chromosomes parents et les autres 50% nouveaux chromosomes sont fabriqués par croisement. Nous avons choisi le croisement a un point dont il s'agit de choisir au hasard un point de croisement pour chaque couple de chromosomes et d'effectuer une permutation des ensembles de gènes se trouvant des deux cotés de ce point des deux parents [54].

La figure 3.19 donne les conséquences de cet opérateur en supposant que les chromosomes 1 et 3, puis 2 et 4 sont appariés et qu'à chaque fois le croisement s'opère avec $P_{crois} = 0.5$.

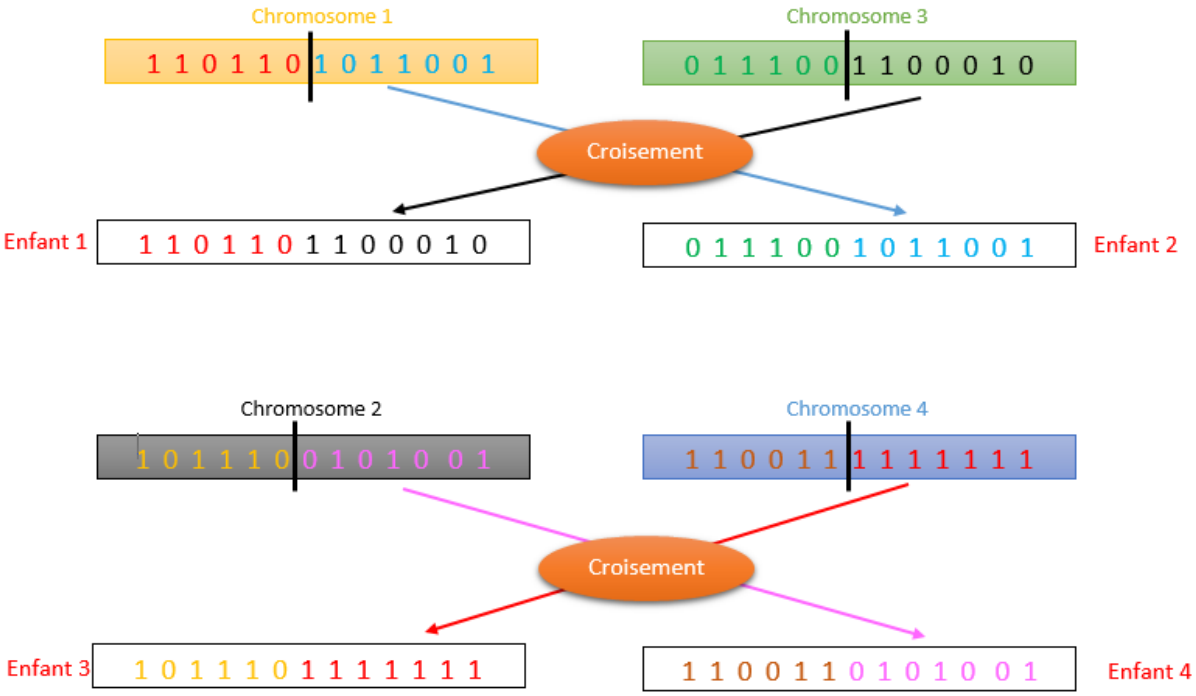

FIGURE 3.19 – Croisement des 4 chromosomes.

3.7.1.6 Étape 6 : Mutation

La mutation est importante pour éviter de tomber dans le problème de l'optimum local, c'est-à-dire éviter de rester « coincé » autour d'une solution pas forcément optimale. Elle consiste à modifier aléatoirement une petite partie de gènes dans certains chromosomes de la nouvelle génération. selon un critère qui est la probabilité de mutation (P_m) qui sert a déterminé la fréquence à laquelle les chromosomes seront mutés [56].
Si $P_m = 1$ cela signifie que tout le chromosome sera changé. Nous avons choisi $P_m = 0.05$ c'est à dire 5% de gènes du chromosome seront inversés. Nous tirons ainsi pour chaque gène(g_i) un chiffre $c(g_i)$ entre 0 et 1 aléatoirement et si ce chiffre est inférieur à P_m on effectue la mutation . Le tableau 3.8 met en évidence ce processus :

Ancien chromosome	**Tirage aléatoire**	**Nouveaux bits**	**Nouveaux chromosomes**
1 1 0 1 **1 0** 1 1 0 0 0 1 0	$c(g_5) = 0.03 < P_m$	**0**	1 1 0 1 **0 1** 1 1 0 0 0 1 0
1 **0** 1 1 1 0 1 1 1 1 1 1 1	$c(g_2) = 0.54 > P_m$	-	1 0 1 1 1 0 1 1 1 1 1 1 1
0 1 1 1 0 **0 1** 0 1 1 0 0 1	$c(g_6) = 0.001 < P_m$	**1**	0 1 1 1 0 **1 0** 0 1 1 0 0 1
1 1 **0** 0 1 1 0 1 0 1 0 0 1	$c(g_3) = 0.062 > P_m$	-	1 1 0 0 1 1 0 1 0 1 0 0 1

TABLE 3.8 – Résultat de la mutation.

Maintenant que la nouvelle population est entièrement crée (voir tableau 3.9 avec max d'accuracy 91,7%, donc nous sommes passés de 89% à 91.7% après une seule génération. nous pouvons de nouveau recommencer la procédure à partir de l'étape de sélection jusqu'à atteindre le critère d'arrêt.

Num	Codage binaire	**Chromosomes**	*Fitness*
1	1 1 0 1 1 0 1 1 0 0 0 1 0	[Age, sex, trestbps, chol, restecg, thalach, ca]	0.821
2	1 0 1 1 1 0 1 1 1 1 1 1 1	[Age, cp, trestbps, chol, restecg, thalach, exang, oldpeak, slope, ca, thal]	0.917
3	0 1 1 1 0 0 1 0 1 1 0 0 1	[sex, cp, trestbps, restecg, exang, oldpeak, thal]	0.875
4	1 1 0 0 1 1 0 1 0 1 0 0 1	[Age, sex, chol, fbs, thalach, oldpeak, thal]	0.849

TABLE 3.9 – Résultats d'évaluation de la nouvelle population.

3.7.1.7 Étape 7 : Critère d'arrêt

Le critère d'arrêt peut prendre la forme suivante :

— Le nombre limité de générations autorisées a été atteint.

— Une stabilité de la population a été atteinte (la population cesse d'évoluer ou n'évolue plus suffisamment) de point de vue de la fonction de fitness.

— Le meilleur compromis dans le cas d'un problème multi-critères a été atteint.

Dans notre étude, nous définissons comme critère d'arrêt le 2ème cas ; c'est à dire jusqu'à atteindre la valeur maximale d'accuracy KNN et les générations ne cessent plus d'évoluer.

3.8 Fractionnement des données

Maintenant que notre ensemble de données est bien préparé nous passons à diviser notre ensemble de données en ensemble d'entraînement que l'on notera (X_train, Y_train), et un ensemble de test, contenant les attributs restants de l'ensemble, que l'on notera (X_test, Y_test). Pour estimer les performances des algorithmes d'apprentissage automatique. Nous avons choisi 80% de l'ensemble de données pour les données d'entraînement et les 20% restants pour les données de test en utilisant la méthode de $train\_test\_split$. Dans ce qui suit, Nous allons présenter les modèles que nous avons choisi pour supporter notre approche de prédiction.

3.9 Classification des données

L'approche de classification est effectuer en 2 façons : La 1ère façon est de classifier les données en utilisant les algorithmes d'apprentissage supervisés KNN, RF, DT, SVM et LR sans sélectionner les attributs pertinents, et la deuxièmes façon est de classifier les données en utilisant les algorithmes cités précédemment avec les attributs sélectionnés par l'algorithme génétique. Dans ce travail on propose le GA-KNN pour prédire précisément si la personne souffre de maladie cardiaque ou non.

Nous sélectionnons les meilleurs hyper-paramètres pour chaque modèle à l'aide de la méthode *grid search cv* (Annexe A section A.3) qui nous permet de tester toutes les combinaisons possibles de paramètres et de comparer les performances pour en déduire le meilleur paramétrage, Le tableau 3.10 montre les hyper-paramètres utilisés dans chaque modèle.

Modèles	**Hyper-paramètres**
LR	'c'=c_space, fit_intercept = Vrai , copy_X = Vrai
SVM	kernel='linéar', gamma=0.01
RF	n_estimators = 500, random_state = 100
DT	random_state=0, max_feature=13
KNN	n_neighbors=4, metric : euclidean

TABLE 3.10 – Les hyper-paramètres utilisés pour chaques algorithmes d'apprentissage supervisé.

3.10 Conclusion

Dans ce chapitre , nous avons présenté des travaux similaires pour la prédiction des maladies cardiaques avec les techniques de l'apprentissage automatiques, puis nous avons défini l'ensemble de données que nous avons utilisé. Aussi, nous avons présenté notre approche proposée pour résoudre la problématique de notre étude avec ses différentes étapes. Dans le chapitre suivant, nous discuterons les résultats obtenus.

Chapitre 4

Résultats et évaluation

4.1 Introduction

Après avoir présenté notre approche et solution proposées dans le chapitre précédent. Nous allons définir dans ce chapitre les différentes bibliothèques et langages de développement utilisés pour le processus d'implémentation, ensuite nous présenterons les résultats des métriques d'évaluation de performances utilisées, enfin nous conclurons ce chapitre avec une comparaison des résultats.

4.2 Environnement de développement

Pour implémenter notre système, nous avons utilisé des plateformes de développement, un langage de programmation et des bibliothèques :

4.2.1 Plateforme de développement

4.2.1.1 Google Colab

Google Colab ou Colaboratory est un service cloud fourni par Google (gratuitement), basé sur Jupyter Notebook et destiné à la formation et à la recherche en machine learning. La plateforme permet de former des modèles de machine learning directement dans le cloud. Il n'est donc pas nécessaire d'installer quoi que ce soit sur notre ordinateur autre qu'un navigateur [57].

4.2.1.2 Jupyter Notebook

Jupyter Notebook est une application Web open source pour créer et partager des documents contenant du code (peut être exécuté directement dans le document), des équations, des images et du texte. A l'aide de cette application, nous pouvons faire du traitement de données, de la modélisation statistique, de la visualisation des données, de l'apprentissage automatique, etc... Elle est disponible par défaut dans la distribution Anaconda [58].

4.2.2 Langage de développement

Nous avons utilisé pour le développement de notre modèle le langage de programmation python.

4.2.2.1 Python

Python est un langage de programmation interprété, multi-paradigme et multiplateformes. Il favorise la programmation impérative structurée, fonctionnelle et orientée objet. Il est doté d'un typage dynamique fort, d'une gestion automatique de la mémoire par ramasse-miettes et d'un système de gestion d'exceptions [59].

4.2.3 Bibliothèques utilisées

4.2.3.1 Matplotlib

Matplotlib est une bibliothèque de traçage disponible pour le langage de programmation Python utilisé pour créer des visualisations statiques, animées et interactives [60].

4.2.3.2 NumPy

NumPy (Numerical Python) est la bibliothèque la plus populaire de calcul scientifique en Python qui permet d'effectuer les calculs scientifiques. Elle propose des fonctions mathématiques complètes, des générateurs de nombres aléatoires, des routines d'algèbre linéaire, des transformées de Fourier ...etc [61].

4.2.3.3 Pandas

Pandas est une bibliothèque open source sous licence BSD qui fournit des structures de données et des outils d'analyse de données hautes performances et faciles à utiliser pour le langage de programmation Python [62].

4.2.3.4 Seaborn

Seaborn est une bibliothèque pour créer des graphiques statistiques en Python. Il est construit sur matplotlib et étroitement intégré aux structures de données pandas. Il fournit une interface avancée pour dessiner des graphiques statistiques attrayants et informatifs [62].

4.2.3.5 Scikit-learn

Scikit-learn est une bibliothèque libre en Python destinée à l'apprentissage automatique. Le grand avantage de scikit-learn est sa courbe d'apprentissage rapide, aussi elle comprend des fonctions pour estimer des régressions logistiques, des algorithmes de classification, et les machines à vecteurs de support. Elle construit sur NumPy, SciPy et matplotlib [63].

4.3 Résultats d'évaluation des performances

L'évaluation permet de mesurer les performances des prédictions par rapport a ses objectifs, notre approche a compris 2 façons : avant et après la sélection des attributs, Nous allons présenter dans ce qui suit les résultats des 2 façons :

4.3.1 Avant la sélection des attributs

Dans le but de savoir à quel point nos algorithmes d'apprentissage automatique sont confus ou qu'ils se trompent. nous avons choisit de les évaluer avec la matrice de confusion, dont elle nous indique le nombre des prédictions correctes et incorrectes pour chaque classe organisées en fonction de la classe prédite. Le figure 4.1 présente les matrices de confusion des différents algorithmes (KNN, RF, DT, SVM et LR) avant la sélection des attributs :

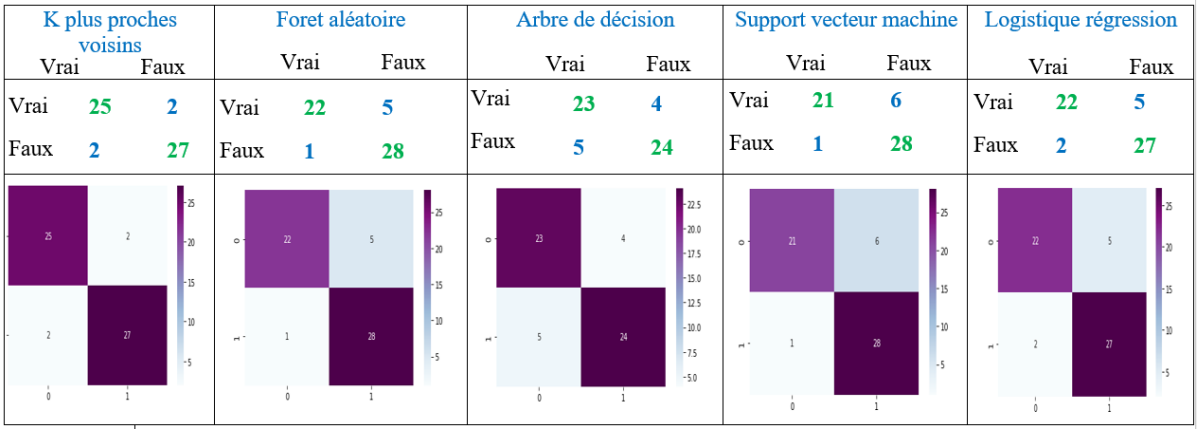

FIGURE 4.1 – Les matrices de confusions des différents algorithmes.

D'après les matrices de confusion des différents algorithmes nous remarquons que le KNN à le taux de vrai positif et négatif (TP + TN = 52) le plus élevé suivi par RF, SVM, LR et DT a le plus bas. ce qui nous montre que KNN peut nous rendre des résultats plus fiables par rapport aux autres modèles, comme notre ensemble de données est équilibré nous prenons l'accuracy comme taux de prédiction de notre étude.

Le tableau 4.1 et la figure 4.2 indiquent les résultats des métriques d'évaluation des performances (Accuracy, Précison, Rappel, F1-score) extraites des matrices de confusion pour chaque algorithmes avant la sélection des attributs :

Modèles	Accuracy	Précision	Rappel	F1-score
RF	89.2%	95.6%	95.6%	87.9%
DT	83.9%	82.1 %	85.1%	83.5%
SVM	87.5%	95.4%	77.7%	85.6%
LR	87.5%	91.6%	81.4%	86.1%
KNN	**92.86**%	92.5%	92.5%	90.9%

TABLE 4.1 – Résultats d'évaluation des performances avant la sélection des attributs.

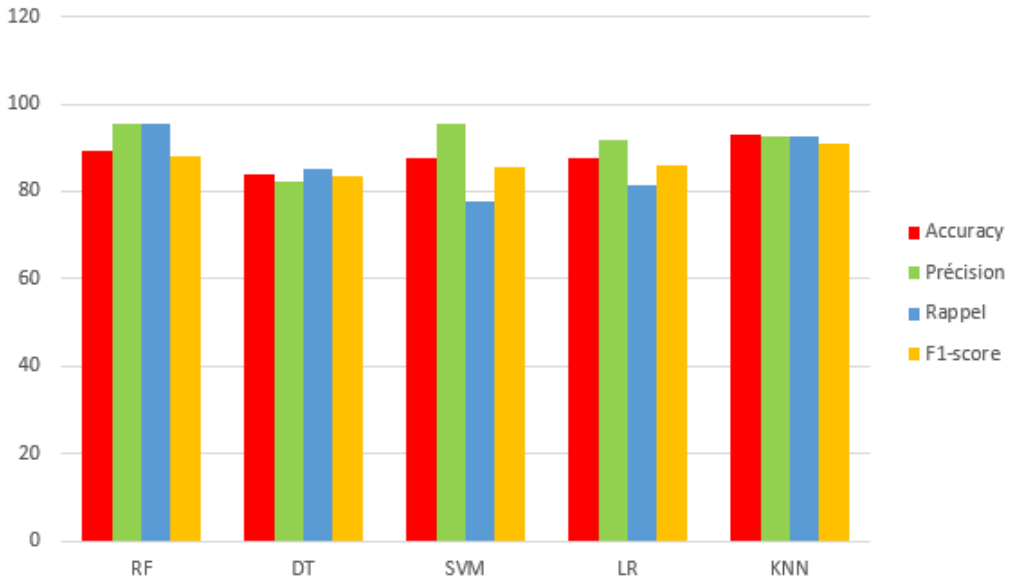

FIGURE 4.2 – Représentation graphique des résultats des métriques d'évaluation avant la sélection des attributs.

D'après les résultats ci dessus on constate que l'algorithme KNN a le taux de prédiction (Accuracy) le plus élevé atteignant 92.86% suivi de de RF 89.2%, aprés SVM et LR avec 87.5%, DT a le taux le plus bas avec 83.9%.

4.3.2 Après la sélection des attributs

Après l'application de l'algorithme génétique notre ensemble de données s'est réduit de 14 à 10 attributs, le tableau 4.2 montre les 3 top résultats de combinaisons d'attributs (individus) sélectionnés selon la fonction d'évaluation fitness.

Nbr d'attributs	Individus	Fonction fitness
9	['age', 'thalach', 'ca', 'cp','exang', 'restecg', 'sex', 'slope','thal']	96.42%
8	['thalach', 'oldpeak', 'ca','cp','exang','sex', 'slope', 'thal']	94.64%
9	['chol', 'thalach', 'ca','cp', 'exang', 'restecg', 'sex', 'slope', 'thal']	92.85%

TABLE 4.2 – Les 3 top meilleurs individus sélectionnés par l'AG.

Nous prenons le meilleur individu sélectionné selon la fonction d'évaluation ['age', 'thalach', 'ca', 'cp','exang', 'restecg', 'sex', 'slope','thal','target'], et on l'applique sur notre ensemble de données pour obtenir un nouveau ensemble de données qui contient que des attributs pertinents afin de retourner des résultats performants.

Notons que la plupart des attributs sélectionnés ont une p-value inférieures à 5% selon les tests statistiques "Anova" et "Khi2" (voir les résultats de l'analyse bivariée section 3.5.2.2 dans chapitre 3) donc nous validons l'hypothèse de qu'ils suggèrent un rôle important dans la prédiction des maladies cardiaques.

Les attributs éliminés par l'algorithme génétique ['fbs','oldpeak','chol','trestbps'], parmi eux 'fbs' et 'trestbps' ont p-value superieur à 5%, nous pouvons dire que le choix d'accuracy KNN comme fonction fitness de AG est important et significatif pour la sélection des attributs.

Le tableau 4.3 englobe les résultats des matrices de confusions des algorithmes d'apprentissage supervisés utilisés après la sélection des attributs.

Modèles	TP	FN	TN	FP
AG-KNN	24	3	28	1
AG-RF	22	1	28	5
AG-DT	23	5	24	4
AG-SVM	21	1	28	6
AG-LR	22	2	27	5

TABLE 4.3 – Résultats des matrices de confusions des modèles utilisés.

Le tableau 4.4 et le graphe illustré dans la figure 4.3 montrent les résultats des métriques d'évaluation des performances des 5 algorithmes appliqués sur le nouveau ensemble de données après la sélection des attributs :

Modèles	Accuracy	Précision	Rappel	F1-score
AG-RF	89.28%	81.48%	95.65%	87.99%
AG-DT	78.57%	82.14%	85.18%	83.63%
AG-SVM	87.5%	77.77%	95.45%	85.7%
AG-LR	87.5%	81.48%	91.66%	86.28%
AG-KNN	**96.42%**	93%	95.2%	94%

TABLE 4.4 – Résultats des métriques d'évaluation des modèles après la sélection des attributs.

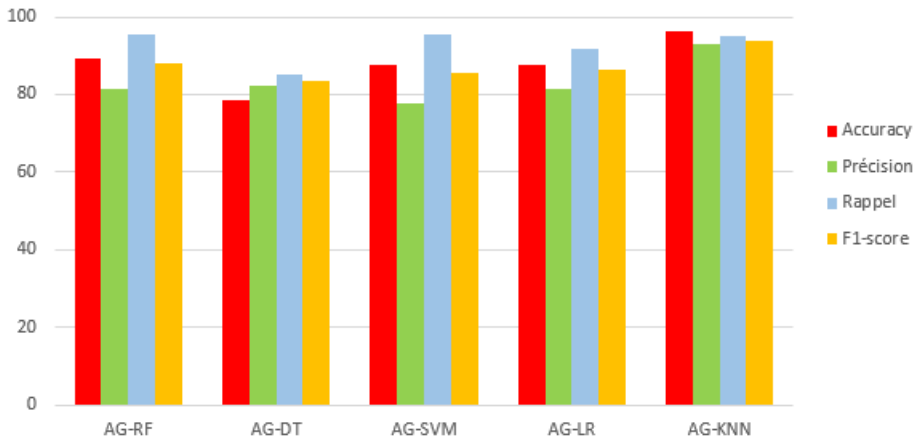

FIGURE 4.3 – Représentation graphique des résultats des métriques d'évaluation après la sélection des attributs.

D'après les résultats des tableaux ci dessus on constate que le modèle AG-KNN a obtenu le meilleur résultat que les autres algorithmes, c'est la conséquence d'élimination de bruit enraciné dans l'ensemble de donnés (élimination des attributs non sélectionnés par l'AG).

4.4 Comparaison des résultats

Le tableau 4.5 et le graphe illustré dans la figure 4.4 représentent la comparaison d'accuracy des algorithmes avant et après la sélection des attributs.

Modèles	Accuracy avec	
	14 attributs	10 attributs
RF	89.2%	89.28%
DT	83.9%	78.57%
LR	87.5%	87.5%
SVM	87.5%	87.5 %
KNN	92.8%	**96.42%**

TABLE 4.5 – Comparaison d'accuracy des algorithmes avant et après la sélection des attributs.

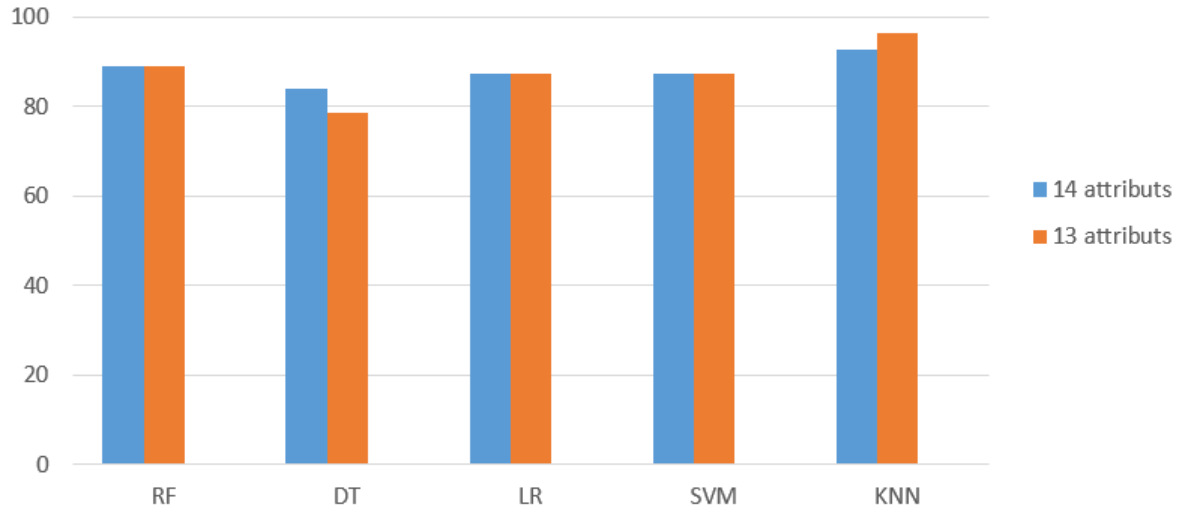

FIGURE 4.4 – Accuracy des modèles avant et après la sélection des attributs.

D'après les résultats ci dessus, nous conclurons les points suivants :

— Le modèle AG-KNN a le taux de prédiction le plus élevé par rapport aux autres algorithmes d'apprentissage supervisés utilisés.

— La sélection des attributs joue un rôle très important dans l'apprentissage automatique, nous remarquons que le taux de KNN à augmenter de 92.86% à 96.42% avec les attributs les plus pertinents, c'est énorme pour un ensemble de données médicale.

— Nous choisissons l'approche AG-KNN comme modèle prédictif de notre problématique vu qu'elle nous a donner le meilleur taux de prédiction atteignant 96.42%.

4.5 Conclusion

Dans ce chapitre, nous avons commencé par recenser les différents environnements et outils de développement utilisés pour l'implémentation de notre approche, ainsi nous avons présenté les résultats d'évaluation des performances des algorithmes utilisés dans notre approche. nous avons concluré ce chapitre avec une comparaison des résultats.

Conclusion générale et perspectives

Les maladies cardiaques peuvent être contrôlées ou leurs effets diminues par la prévention ou la gestion des facteurs de risque. Il est difficile de déterminer manuellement les chances de contracter une maladie cardiaque en fonction des facteurs de risque, c'est pour cela l'apprentissage automatique s'avère efficace pour aider à prendre des décisions et des prévisions à partir de la grande quantité de données produites par le secteur de santé, L'idée principale est de créer des systèmes prédictifs qui visent à améliorer la qualité des soins aux patients et la prédiction de la maladie.

L'objectif de notre travail, était la mise en place d'un système prédictif pour les maladies cardiaques, et pour atteindre cet objectif, nous avons utilisé l'ensemble de données Cleaveland heart dataset qui est un ensemble de données médicale dont il englobe les risques des maladies cardiaques, nous avons bien analysé et pré-traité cet ensemble de données afin de lui rend prêt pour la phase de classification, nous avons l'entraîner en deux façons : La 1ère façon est de classifier les données en utilisant les algorithmes d'apprentissage supervisés KNN, RF, DT, SVM et LR sans sélectionner les attributs pertinents, et la deuxièmes façon est de classifier les données en utilisant les algorithmes cités précédemment avec les attributs sélectionnés par l'algorithme génétique. Dans ce travail on propose le GA-KNN pour prédire précisément si la personne souffre de maladie cardiaque ou non.

Nous pouvons conclure que la sélection des attributs avec l'algorithme génétique a non seulement réduit la taille de l'ensemble de données de 14 attributs à 10 attributs pertinents ce qui implique un temps de réponse réduit, et une élimination de bruit dans l'ensemble de données ce qui fait une prédiction plus performante et plus fiable, mais aussi a amélioré le

rendement de taux de prédiction des algorithmes d'apprentissage automatiques supervisés dont KNN à atteint 96.42% de taux de prédiction. c'est pour cela l'approche AG-KNN est choisit comme modèle prédictif de notre système.

Avant d'aller aux perspectives, ce projet a été bénéfique à plusieurs niveaux :

- Au niveau technique, on a eu l'occasion d'enrichir nos connaissances concernant les outils et environnements de développement tels que Jupyter Notbook, et Google Colaboratory. Ainsi, on a eu l'opportunité de maîtriser le langage de programmation Python et ses bibliothèques.
- Au niveau personnel, cette expérience nous a permis de découvrir le milieu de la science des données avec tout ce qu'il exige de règles, d'analyse et la mise à jour de ses données ce qui va nous aider à s'orienter vers le choix du domaine professionnel dans des meilleurs niveaux d'appartenance et savoir.

Comme perspectives :

- Nous souhaitons combiner d'autres classificateurs d'apprentissage approfondie (deep learning) avec l'algorithme génétique pour avoir des meilleurs résultats de prédictions des maladies cardiaques.
- Nous envisageons d'améliorer notre approche avec l'intégration des variétés de KNN comme le KNN pondéré dans la fonction fitness de l'algorithme génétique afin de pondérer les attributs avant de les sélectionner ce qui fournit une sélection plus pertinente d'attributs.
- Effectuer une analyse sur notre étude avec d'autres ensembles de données ou des données extraites des experts médicaux afin d'inclure de nouvels attributs pour tester les performances du système proposés.

Bibliographie

[1] Lolita p, pharm d, structure anatomique du cœur. https://www.sante-sur-le-net.com/maladies/cardiologie/generalites-coeur/. Consulté le 23 fev 2022.

[2] H mayar, comment fonctionne le cœur, https ://www.coeuretavc.ca/maladies-du-coeur/qu-est-ce-que-les-maladies-du-coeur/comment-fonctionne-le-coeur. https://www.coeuretavc.ca/maladies-du-coeur/qu-est-ce-que-les-maladies-du-coeur. Consulté le 23 fev 2022.

[3] Benahmed, selim, beutler, laurence, et chabloz, boris. l'éducation thérapeutique post-infarctus du myocarde : le défi de la prévention tertiaire . 2015. thèse de doctorat. haute ecole arc santé.

[4] Santé, accident vasculaire cérébral (avc) : ce qu'il faut savoir. https://www.santemagazine.fr/sante/fiche-maladie/avc-accident-vasculaire-cerebral-177349. Consulté le 17-05-2022.

[5] Dremstime.s, figure 1, consulté le 17-05-2022. https://fr.dreamstime.com/images-libres-droits-ath%C3%A9roscl%C3%A9rose-art%C3%A9rioscl\%C3%A9rose-image31746909. Consulté le 17-05-2022.

[6] Pr.chetibi, angine de poitrine : 30 à 40% des jeunes sujets ne consultent pas en cas de douleurs. https://www.aps.dz/sante-science-technologie/89248-angine-de-poitrine-30-a-40-des-jeunes-sujets-ne-consultent\-pas-en-cas-de-douleurs. Publié Le : Jeudi, 09 Mai 2019 16 :42.

[7] Sympa, qu'est-ce que l'insuffisance cardiaque et pourquoi est-il important que tout le monde connaisse cette affection ? https://sympa-sympa.com/creation-bien-etre/

\quest-ce-que-linsuffisance-cardiaque-et-pourquoi\
-est-il-important-que-tout-le-monde-connaisse-\cette-affection.
Consulté le 17-05-2022.

[8] Zakia messaouidi, spiria, blogue et discussions, les 3 étapes essentielles de l'apprentissage automatique (machine learning), 22 janvier 2020, consulté le 24/05/2022. https://www.spiria.com/fr/blogue/intelligence-artificielle/3-etapes-essentielles-apprentissage\-automatique-machine-learning/.

[9] Bastien maurice, les différents types d'apprentissage, 15 septembre 2018, consulté le 05/05/2022. https://deeplylearning.fr/cours-theoriques%-deep-learning/les-differents-types-dapprentissage/.

[10] Halliche amel, classification supervisée à la base de knn avec pondération d'attributs par l'algorithme génétique, thèse doctorat, usthb, 2015.

[11] Merabti, youcef. optimisation des réseaux de neurones mlp par l'algorithme hybride ag-rt pour le contrôle d'un système non linéaire. 2015.

[12] Bel hadj ali, nizar. etude de la conception globale des structures en construction métallique. optimisation par les algorithmes génériques. 2003. thèse de doctorat. chambéry.

[13] Ghuiremekinzel, matrice de confusion. https://www.kaggle.com/questions-and-answers/118209?fbclid=IwAR2n0BwOfhJ10szMCbmk2lacPOcCdREaE3xm-myTN7YUmCmqloKH9aR8tII.
Consulté le 04 mai 2022.

[14] Chui, kwok tai, alhalabi, wadee, pang, sally shuk han, et al. disease diagnosis in smart healthcare : Innovation, technologies and applications. sustainability, 2017, vol. 9, no 12, p. 2309.

[15] Gaut, cours de physiologie, le coeur. http://f2.quomodo.com/5C852034/uploads/8489/physiologie.

[16] Ornstein, steven, jenkins, ruth g., nietert, paul j., et al. une intervention d'amélioration de la qualité multi-méthodes pour améliorer les soins cardiovasculaires préventifs : un essai randomisé en grappes. annales de médecine interne , 2004, vol. 141, n° 7, p. 523-532.

[17] Oms, mvc. https://www.iaea.org/fr/themes/les-maladies-cardiovasculaires. Consulté le 17-05-2022.

[18] Thomas cascino.md, university of michigan, le manuelle msd le grand publique, dernière révision totale, juil 2021| dernière modification du contenu juil 2021. https://www.msdmanuals.com/fr/accueil/troubles-cardiaques-et-vasculaires/diagnostic-des-maladies-cardiovasculaires/introduction-au-diagnostic-des-maladie Consulté le 17-05-2022.

[19] Lachance, kim. facteurs de risque d'insuffisance rénale chronique chez les greffés cardiaques : du phénotype aux tests pharmacogénomiques. 2015.

[20] Laroche, jean-pierre, miserey, gilles, guilbert, bruno, et al. medecine vasculaire. concours medical, 2005, vol. 127, no 35, p. 3-4. consulté 30/03/2022.

[21] Janin, emilie. participation du pharmacien d'officine à l'éducation thérapeutique des patients après un programme de réadaptation cardiaque : pérennisation de l'observance médicamenteuse, du respect des règles hygiéno-diététiques et du maintien de l'activité physique. 2015. thèse de doctorat. université de lorraine.

[22] Nguyen, cam linh. prédiction de la réponse aux traitements de vivo de tumeurs basées sur le profil moléculaire des tumeurs par apprentissage automatique, 2019. thèse de doctorat. aix-marseille.

[23] Mifdal, r. (2019). application des techniques d'apprentissage automatique pour la prédiction de la tendance des titres financiers (doctoral dissertation, École de technologie supérieure).

[24] Bellahmer, hacene. implémentation et évaluation d'un modèle d'apprentissage automatique pour l'estimation de la valeur marchande de propriétés immobilières. 2020. thèse de doctorat. université mouloud mammeri.

[25] Desjardins, julie. l'analyse de régression logistique. tutorial in quantitative methods for psychology, 2005, vol. 1, no 1, p. 35-41.

[26] La rédaction jdn, machine à vecteurs de support (svm) : Définition et cas d'usage. https://www.journaldunet.fr/web-tech/guide-de-l-intelligence-artificielle\/1501879-machine-a-vecteurs-de-support-svm/. Consulté le 18-05-2022.

[27] Mhennaoui abdelghani djouadi cherif, la prédiction des maladies cardiaques à l'aide des techniques d'apprentissage automatique, mémoire de master, uamo bouira, 2020.

[28] Bouamra amira, allalou nour el houda, vers une approche de recommandation de services sensible au contexte utilisant les techniques de machine learning, memoire de master, uamo bouira, 2021.

[29] Chamroukhi, faïcel. classification supervisée : Les k-plus proches voisins. mémoire de fin d'étude, université du sud toulon–var, 2013.

[30] Ghoulam, aicha, barigou, fatiha, belalem, ghalem, et al. using local grammar for entity extraction from clinical reports. 2015.

[31] DescÔteaux, steve. les règles d'association maximale au service de l'interprétation des résultats de la classification. 2014. thèse de doctorat. université du québec à trois-rivières.

[32] Labiad, ali. sélection des mots clés basée sur la classification et l'extraction des règles d'association. 2017. thèse de doctorat. université du québec à trois-rivières.

[33] Mr.mint, 2018 consulté le 22 avril 2022, https ://mrmint.fr/ « mr.mint machine learning made easy.

[34] Sabrina. a , apprentissage automatique et réduction du nombre de dimensions. `http://www-ia.lip6.fr/tollaris/articles/these/node7.html`. 2008 Consulté le 22 avril 2022.

[35] Lavanya devi, n. et thirumurugan, p. classification du cancer du col de l'utérus à partir d'images de frottis de pap à l'aide de moyennes c floues modifiées, pca et knn. journal de recherche de l'iete , 2021, p. 1-8.

[36] Wikipedia, k-nearest neighbors algorithm. `<https://en.wikipedia.org/wiki/knearest_neighbors_algorithm`. 2016 Consulté le 21 avril 2022.

[37] Hilali, hassane. application de la classification textuelle pour l'extraction des règles d'association maximales. 2009. thèse de doctorat. université du québec à trois-rivières.

[38] Challita, nicole. contributions à la sélection des attributs de signaux non stationnaires pour la classification. 2018. thèse de doctorat. université de technologie de troyes.

[39] El akadi, ali. contribution à la sélection de variables pertinentes en classification supervisée : Application à la sélection des gènes pour les puces à adn et des caractéristiques faciales. 2012.

[40] Gabriel comier . l'algorithme genetique en genie electrique : application a l'ellipsometrie et aux reseaux de bragg. 2006.

[41] Yachba, khadidja, gelareh, shahin, et bouamrane, karim. gestion du stockage des conteneurs dangereux à l'aide de l'algorithme génétique. transports et télécommunications , 2016, vol. 17, n° 4, p. 371.

[42] VallÉe, thomas et yildizoĞlu, murat. présentation des algorithmes génétiques et de leurs applications en économie. revue d'économie politique, 2004, p. 711-745.

[43] Myriam emilion - directrice marketing à jedha, matrice de confusion : comment la lire et l'interpréter ?, https ://www.jedha.co/formation-ia/matrice-confusion , consulté le 11/07/2022.

[44] Sharma, vineet, rasool, akhtar, et hajela, gaurav. prédiction des maladies cardiaques à l'aide de dnn. dans : 2020 second international conference on inventive research in computing applications (icirca) . ieee, 2020. p. 554-562.

[45] Nahiduzzaman, md, nayeem, md julker, ahmed, md toukir, et al. prédiction des maladies cardiaques à l'aide d'un réseau de neurones perceptrons multicouches et d'une machine à vecteurs de support. en : 2019 4ème conférence internationale sur les technologies électriques de l'information et de la communication (eict) . ieee, 2019. p. 1-6.

[46] El hamdaoui, halima, boujraf, saïd, chaoui, nour el houda, et al. l'invention concerne un système d'assistance clinique pour la prédiction des maladies cardiaques à l'aide de techniques d'apprentissage automatique. dans : 2020 5th international conference on advanced technologies for signal and image processing (atsip) . ieee, 2020. p. 1-5.,.

[47] Sujatha, p. et mahalakshmi, k. Évaluation des performances d'algorithmes d'apprentissage automatique supervisé dans la prédiction des maladies cardiaques. dans : Conférence internationale ieee 2020 pour l'innovation technologique (inocon) . ieee, 2020. p. 1-7.

[48] Singh, archana et kumar, rakesh. prédiction des maladies cardiaques à l'aide d'algorithmes d'apprentissage automatique. dans : Conférence internationale 2020 sur le génie électrique et électronique (ice3) . ieee, 2020. p. 452-457.

[49] Sateesh kumar, r. et sameen fatima, s. heart disease prediction using extended knn (e-knn). in : Smart computing techniques and applications. springer, singapore, 2021. p. 565-572. issn : 2277-3878.

[50] Sateesh kumar, r. et sameen fatima, s. heart disease prediction using extended knn (e-knn). in : Smart computing techniques and applications. springer, singapore, 2021. p. 565-572.

[51] Analyse de données avec spss, carricano, manu and poujol, fanny and bertrandias, laurent, 2010, pearson education france.

[52] Falissard, b. déploiement d'une matrice de corrélation sur la sphère unité de r 3. revue de statistique appliquée, 1995, vol. 43, no 2, p. 35-48.

[53] Wikipédia, les valeures abberantes. `https://fr.wikipedia.org/wiki/Donn%C3%A9e_aberrantetext=En%20statistique%2C%20une%20donn%C3%A9e%20aberrante,les%20valeurs%20%C2%AB%20normalement%20%C2%BB%20mesur%C3%A9es.` Consulté le 12/05/2022.

[54] Saad, ihsen, tangour, fatma, et borne, pierre. application des algorithmes génétiques aux problèmes d'optimisation : Métaheuristiques pour l'optimisation difficile. ree. revue de l'électricité et de l'électronique, 2009, no 4.

[55] Horn, jeffrey, nafpliotis, nicholas, et goldberg, david e. un algorithme génétique de pareto niché pour l'optimisation multiobjectif. in : Actes de la première conférence ieee sur le calcul évolutionnaire. congrès mondial ieee sur l'intelligence computationnelle . iee, 1994. p. 82-87.

[56] Ghezali, yamina. le problème de la tournée de véhicule avec contrainte de capacité par l'algorithme génétique. 2021. thèse de doctorat. université de bordj bou arreridj faculty of mathematics and computer science.

[57] Alves, francisco regis vieira et vieira, renata passos machado. the newton fractal's leonardo sequence study with the google colab. international electronic journal of mathematics education, 2019, vol. 15, no 2, p. em0575.

[58] Randles, bernadette m., pasquetto, irene v., golshan, milena s., et al. using the jupyter notebook as a tool for open science : An empirical study. in : 2017 acm/ieee joint conference on digital libraries (jcdl). ieee, 2017. p. 1-2.

[59] Hellmann, doug. la bibliothèque standard python par exemple . upper saddle river, États-unis : Addison-wesley, 2011.

[60] Hunter, john d. matplotlib : A 2d graphics environment. computing in science engineering, 2007, vol. 9, no 03, p. 90-95.

[61] Oliphant, travis e. a guide to numpy. usa : Trelgol publishing, 2006.

[62] Kumar, arun et panda, supriya p. une enquête : comment python pitche dans le it-world. dans : 2019 international conference on machine learning, big data, cloud and parallel computing (comitcon) . ieee, 2019. p. 248-251.

[63] Kramer, olivier. scikit-apprendre. dans : Apprentissage automatique pour les stratégies d'évolution . springer, cham, 2016. p. 45-53.

[64] pr thomas , dr jean-philippe, rivière directeur médical de doctissimo, 10 conseils pour prévenir les maladies cardiovasculaires, hôpital la pitié-salpêtrière à paris, mis à jour le 13-06-2019 à 11h52.

[65] Morand, Élisabeth. data science : fondamentaux et études de cas, machine learning avec python et r par eric biernat et michel lutz. population, édition anglaise , 2018, vol. 73, n° 2, p. 386-387.

[66] Berhoum adel, belhadi mohamed tahar, apprentissage automatiquedes maladies cardiaques dans les systemes big data, universite echahid hamma lakhdar - el oued, 22 juin 2019.

[67] M. han and h. zhang, business intelligence architecture based on internet of things, journal of theoretical and applied information technology, 2013, vol 50, page 90-95.

[68] Lo, aw, mamaysky, h., wang, j. (2000). fondements de l'analyse technique : algorithmes informatiques, inférence statistique et mise en œuvre empirique. le journal des finances , 55 (4), 1705-1765.

[69] Settouti, nesma et hafa, amel. approche filtre pour la sélection des gènes pertinents des données biopuces du cancer du côlon. 2013.

[70] Chouaib, hassan. sélection de caractéristiques : méthodes et applications. université paris descartes : Paris, france , 2011.

[71] Dash, manoranjan et liu, huan. sélection des fonctionnalités pour la classification. analyse intelligente des données , 1997, vol. 1, n° 1-4, p. 131-156.

[72] Charik, khalissa et charik, loubna. approche filtre par la sélection de données multi-sensorielles pour l'aide au diagnostic médical. 2020. thèse de doctorat. univ m'sila.

[73] Ouanas, houdham, kouidri, yacine, et al. classification de la maladie d'alzheimer à l'aide de l'apprentissage statistique. 2021. thèse de doctorat. university of m'sila.

[74] John, george h., kohavi, ron, et pfleger, karl. caractéristiques non pertinentes et problème de sélection de sous-ensemble. dans : Procédures d'apprentissage automatique 1994 . morgan kaufmann, 1994. p. 121-129.

[75] Chandrashekar, girish et sahin, ferat. une enquête sur les méthodes de sélection des fonctionnalités. informatique et génie électrique , 2014, vol. 40, n° 1, p. 16-28.

[76] Ouanas, houdham, kouidri, yacine, et al. classification de la maladie d'alzheimer à l'aide de l'apprentissage statistique. 2021. thèse de doctorat. university of m'sila.

[77] Michalewicz, zbigniew. ga : pourquoi fonctionnent-ils ?. dans : Algorithmes génétiques + structures de données = programmes d'évolution . springer, berlin, heidelberg, 1996. p. 45-55.

[78] Bastien l, machine learning : Définition, fonctionnement, utilisations. `https://datascientest.com/machine-learning-tout-savoir`, note= Consulté le 04 mai 2022,.

[79] Hajar, rachel. facteurs de risque de maladie coronarienne : perspectives historiques. vues du cœur : le journal officiel de la gulf heart association , 2017, vol. 18, n° 3, p. 109.

[80] World health organization, et al. la readaptation des maladies cardio-vasculaires : rapport d'un comite d'experts de l'oms [reuni a geneve du 23 au 29 juillet 1963]. 1964.

[81] Rahmat, dadi, putra, andika a., setiawan, agung w., et al. prédiction des maladies cardiaques à l'aide de k-nearest neighbor. dans : 2021 international conference on electrical engineering and informatics (iceei) . ieee, 2021. p. 1-6.

[82] Mackay, judith, mensah, george a., et greenlund, kurt. l'atlas des maladies cardiaques et des accidents vasculaires cérébraux . organisation mondiale de la santé, 2004.

[83] Ambekar, sayali et phalnikar, rashmi. prédiction du risque de maladie à l'aide d'un réseau neuronal convolutif. en : 2018 quatrième conférence internationale sur le contrôle et l'automatisation des communications informatiques (iccubea) . ieee, 2018. p. 1-5.

[84] Latha, c. beulah christalin et jeeva, s. carolin. amélioration de la précision de la prédiction du risque de maladie cardiaque basée sur des techniques de classification d'ensemble. l'informatique en médecine déverrouillée , 2019, vol. 16, p. 100203.

[85] Gavhane, aditi, kokkula, gouthami, pandya, isha, et al. prédiction des maladies cardiaques à l'aide de l'apprentissage automatique. dans : 2018 deuxième conférence internationale sur l'électronique, la communication et les technologies aérospatiales (iceca) . ieee, 2018. p. 1275-1278.

[86] H motada et h liu, feature slection extraction and constructio, the handbook of datamining, 2003, lawrence erlbaum associate, 409-423.

[87] BÄrecke, thomas, lesot, marie-jeanne, akdag, herman, et al. stratégie de fusion d'informations exploitant le réseau des sources. 8ème atelier sur la fouille de données complexes complexité liée aux données multiples, 2011.

[88] Haliche, amel. classification supervisée à base de knn avec pondération d'attributs par l'algorithme génétique. 2015. thèse de doctorat. faculté d'electronique et d'informatique.

[89] Imen trabelsi, prédiction d'obsolescence basée sur la sélection conjointe de caractéristiques et les techniques d'apprentissage automatique, 2009.

[90] Settouti, nesma et hafa, amel. approche filtre pour la sélection des gènes pertinents des données biopuces du cancer du côlon. 2013. ref ffhal-00843080f.

[91] Mr bekhti mohammed anès, la sélection de variables pour la reconnaissance du diabète en utilisant une approche neuronale. 2011-2012. thèse de master, université tlemcen,.

[92] Deekshatulu, bl, chandra, priti, et al. classification des maladies cardiaques à l'aide du k-plus proche voisin et d'un algorithme génétique. technologie procedia , 2013, vol. 10, p. 85-94.

[93] Dahan, marie-line. l'effet du vieillissement sur la microcirculation cutanée. 2008. thèse de doctorat. université claude bernard-lyon i.

[94] Ayoub chebbi, data science : Les bibliothèques de python, jan 14, 2019. https://medium.com/ayoubchebbi/data-science-les-biblioth%C3%A8ques-de-python-51952ca40dd6. Consulté le 11/05/2022.

[95] Databird. tout savoir sur le langage de programmation python. https://www.data-bird.co/python/langage-python. Consulté le 11/05/2022.

[96] Le figaro, hypertension artérielle : de quoi s'agit-il ?, howpublished = https://institut.amelis-services.com/sante/hypertension/hypertension-arterielle-chez-les-personnes-agees/,, note=Consulté le 17-05-2022,.

[97] Kotsiantis, sotiris b., zaharakis, ioannis d., et pintelas, panayiotis e. machine learning : a review of classification and combination techniques. revue de l'intelligence artificielle , 2006, vol. 26, n° 3, p. 159-190.

[98] Sun, shiliang et huang, rongqing. un algorithme adaptatif des k plus proches voisins. en : 2010 septième conférence internationale sur les systèmes flous et la découverte des connaissances . ieee, 2010. p. 91-94.

[99] Pan, zhibin, wang, yikun, et pan, yiwei. a new locally adaptive k-nearest neighbor algorithm based on discrimination class. knowledge-based systems, 2020, vol. 204, p. 106185.

[100] Cherif, walid. optimisation de l'algorithme k-nn par clustering et coefficients de fiabilité : application au diagnostic du cancer du sein. procedia informatique , 2018, vol. 127, p. 293-299.

[101] Uddin, shahadat, haque, ibtisham, lu, haohui, et al. analyse comparative des performances de l'algorithme k-plus proche voisin (knn) et de ses différentes variantes pour la prédiction des maladies. rapports scientifiques , 2022, vol. 12, n° 1, p. 1-11.

[102] Han.hs, karypis.g kumar, lors de la conférence asie-pacifique sur la découverte des connaissances et l'exploration de données, 53–65 (springer), conférence, consulté le 17-05-2022.

[103] Yigit, halil. une approche de pondération pour le classificateur knn. dans : Conférence internationale 2013 sur l'électronique, l'informatique et le calcul (icecco) . ieee, 2013. p. 228-231.

[104] Dhar, joydip, shukla, ashaya, kumar, mukul, et al. un k-plus proche voisin pondéré pour l'extraction de classification. arxiv preprint arxiv :2005.08640 , 2020.

[105] Keller.jm, gray.mr givens.ja, un algorithme flou k-plus proche voisin, ieee trans, syst, homme cybern, vol 15, page 580-585, 1985.

[106] Alkasassbeh, mouhammd, altarawneh, ghada a., et hassanat, ahmad. sur l'amélioration des performances des classificateurs voisins les plus proches à l'aide de la métrique de distance hassanat. arxiv preprint arxiv :1501.00687 , 2015.

[107] Gou, jianping, du, lan, zhang, yuhong, et al. un nouveau classificateur k-plus proche voisin pondéré en fonction de la distance. j. inf. calcul. sci , 2012, vol. 9, n° 6, p. 1429-1436.

[108] Uddin, shahadat, haque, ibtisham, lu, haohui, et al. analyse comparative des performances de l'algorithme k-plus proche voisin (knn) et de ses différentes variantes pour la prédiction des maladies. rapports scientifiques , 2022, vol. 12, n° 1, p. 1-11.

[109] Karine levy-heidmann, en médecine, les impacts réels de l'intelligence artificielle. `https://www.ouest-france.fr/sante/\en-medecine-les-impacts-reels-de-l-intelligence-artificielle-5449707`. 17-12-2017 Consulté le 17-05-2022.

[110] Branch and bound, wikipedia, https ://en.wikipedia.org/wiki/branch_and_bound. `https://en.wikipedia.org/wiki/Branch_and_bound`. Consulté le 17-05-2022.

[111] Yacoub, méziane et bennani, younès. détection et sélection d'informations pertinentes : application à la reconnaissance de visages. in : Content-based multimedia information access-volume 1. 2000. p. 649-664.

[112] GuÉrif, sébastien et bennani, younès. sélection de variables en apprentissage numérique non supervisé.univ. paris , 2007.

[113] Dy, jennifer g. et brodley, carla e. feature subset selection and order identification for unsupervised learning. dans : Icml . 2000. p. 247-254.

[114] Murphy, thomas brendan, dean, nema, et raftery, adrian e. variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications. the annals of applied statistics, 2010, vol. 4, no 1, p. 396.

[115] Hong, yi, kwong, sam, chang, yuchou, et al. unsupervised feature selection using clustering ensembles and population based incremental learning algorithm. pattern recognition, 2008, vol. 41, no 9, p. 2742-2756.

[116] Guyon, isabelle et elisseeff, andré. une introduction à la sélection de variables et de fonctionnalités. journal de recherche sur l'apprentissage automatique , 2003, vol. 3, no mars, p. 1157-1182.

[117] Koller, daphné et sahami, mehran. vers une sélection optimale des fonctionnalités . laboratoire d'information de stanford, 1996.

[118] Bendana, rokia. sélection d'attributs basée sur un algorithme génétique neuronal.2007.

[119] Setu kumar basak, comment effectuer une sélection basée sur la roulette et le rang dans un algorithme génétique ?, 5-07-2018. `https://setu677.medium.com/how-to-perform-\roulette-wheel-and-rank-based-selection-in-a-\genetic-algorithm-d0829a37a189`. Consulté le 17-05-2022.

[120] Sahel africain au gilbert vedrenne professeur, exemple de sélection par roulette biaisée pour 4 individus. `https://www.researchgate.net/figure/Figure-E1-Exemple-de-selection-par-roulette-biaisee-pour-4-individus-K-i-avec-J` `fig35_228699661`. Consulté le 17-05-2022.

[121] Latex. `https://fr.wikipedia.org/wiki/LaTeX,Consultéle24/06/2022`. Consulté le 24/06/2022.

[122] Grid search cv , scikit-learn , https ://scikit-learn.org/stable/modules/generated/sklearn.model_selection.gridsearchcv.html. Consulté le 24/06/2022.

Annexe A

Annexe

A.1 Tests statistiques

A.1.1 Test de khi2 pearson (test d'indépendance)

Le test d'indépendance du khi-carré (l'écriture anglaise est « chi-square ») a été développé par Karl PEARSON (1857-1936), est un test non paramétrique permettant de vérifier s'il existe un lien entre deux caractères issu d'une même population donnée lorsque ces caractères sont qualitatifs où lorsque'un caractère est quantitatif et l'autre qualitatif, ou bien encore lorsque les deux caractères sont quantitatifs mais que les valeurs ont été regroupées.[67].

Les hypothèses statistiques de khi2 se représentent comme suite : Soit$Y$ la variable dépendante et X la variable indépendante.

— $H0$: les variables$X$ et Y sont indépendantes si *p-value* $> 5\%$;

— $H1$: les variables X et$Y$ sont dépendantes l'une de l'autre si *p-value* $< 5\%$.[8]

A.1.2 Test d'Anova (Analyse de variance)

L'analyse de variance (ANOVA) en anglais (Analysis of Variance), est un outil statistique créé par Ronald Fisher [9]. Il est connu sous le nom d'analyse de la variance de Fisherest est un test statistique permet d'évaluer l'influence d'une ou plusieurs variables indépendantes catégorielles sur une variable dépendante continue [8].

Les hypothèses statistiques pour la réalisation du test d'ANOVA sont :

— $H0$: les groupes proviennent de la même population et leurs moyennes sont toutes

égales, ce qui implique que la "cible" et une variable indépendante ne sont PAS corrélées (l'hypothèse(H0) est acceptée que si $p - value < 0,05$).

— $H1$: les moyennes ne sont pas toutes égales.

A.2 Outils de rédaction

Nous avons utilisé pour la rédaction de ce mémoire le langage Latex et l'éditeur Overleaf.

A.2.1 Latex

Latex est un langage et un système de composition de documents. Il s'agit d'une collection de macro-commandes destinées à faciliter l'utilisation du « processeur de texte ». LaTeX permet de rédiger des documents dont la mise en page est réalisée automatiquement en se conformant du mieux possible à des normes typographiques. Une fonctionnalité distinctive de LaTeX est son mode mathématique, qui permet de composer des formules complexes [121].

A.2.2 Overleaf

Overleaf est un éditeur LaTeX en ligne, collaboratif en temps réel.

A.3 Grid search cv

Grid search cv c'est méthode de recherche exhaustive sur les valeurs de paramètre spécifiées pour un estimateur, en implémentant une méthode "fit" et une méthode "score". Elle implémente également "score_samples", "predict", "predict_proba", "decision_function", "transform" et "inverse_transform" s'ils sont implémentés dans l'estimateur utilisé. Les paramètres de l'estimateur utilisé pour appliquer ces méthodes sont optimisés par une recherche de grille à validation croisée sur une grille de paramètres [122].