


République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Akli Mohand Oulhadj de Bouira

Faculté des Sciences et des Sciences Appliquées

Département d'Informatique

Mémoire de Master

en Informatique

Spécialité : *Génie des systèmes d'informatique*

Thème

La prédiction des maladies cardiaques à l'aide des techniques d'apprentissage automatique

---

Encadré par

— Brahimi Farida

Réalisé par

— Mehennaoui Abdelghani

— Djouadi Chérif

2020/2021

Remerciements

Tout d'abord, nous tenons à remercier Dieu le Tout-Puissant pour ses bénédictions et sa miséricorde envers nous pendant notre étude et l'achèvement de ce projet. Et que les prières et la paix de Dieu soient sur le sceau de son prophète Mohammed, que les prières et la paix de Dieu soient sur lui, sa famille et ses compagnons.

Notre gratitude et notre appréciation vont également à tous ceux qui nous ont soutenus, famille et amis. Merci pour votre enthousiasme et vos encouragements. C'est une expérience humiliante de reconnaître ceux qui ont, la plupart du temps par gentillesse, aidé tout au long de notre voyage, même si c'était avec de simples mots gentils ou un petit conseil.

Nous devons notre plus profonde gratitude à notre encadreur Mme F.Brahimi pour son aide et son soutien fréquents, pour son temps, sa patience et ses conseils, il n'aurait pas été possible de mener à bien ce projet sans son aide, nous guidant dans un domaine d'étude et un domaine nouveau pour nous. Nous avons définitivement gagné et appris beaucoup de ce voyage grâce à elle.

Dédicaces

Je dédie mon travail
A mon pére et a ma mére, et mon frére.

Pour mes amis, qui mon soutenu durant cette épreuve, ainsi que mon collegue dans ce projet Abdelghani.

Djouadi Chérif.

Dédicaces

Je dédie mon travail
A ma grande famille bien aimante.

A mes amis qui m'on bien aidé et soutenu durant toute ces ans d'étude, ainsi que mes collegues de master 2 GSI, son oublier mon collegue de project Djouadi cherif.

Abdelghani Mehennaoui

ملخص

يعد الكشف المبكر من بين المل المهمة التي تصنع فرق في علاج الأمراض ، ولهذا يشهد العالم في السنوات الأخيرة ارتباط و انتشار واسع لمناهج التعلم الآلي و التعلم العميق في التنبؤ بالأمراض للأشخاص قبل حدوثها أو حتى الكشف عنها ، وذلك لتقليل الخطر على حياة الإنسان ، من بين الأمراض التي تشكل خطر كبير على حياة الإنسان هي أمراض القلب لذلك لا بد من التنبؤ بِها و معالجتها قبل تطورها الهدف من عملنا هو محاولة التنبؤ المبكرة بأمراض القلب و ذلك باستعمال نهج التعلم الآلي في مجال علم الأمراض للقيام بذلك ، اخترنا استخدام بعض خوارزميات التعلم الآلي ، و ذلك لاختيار الخوارزمية التي كانت نسبت تنبؤها عالية بعد تدريبها على مجموعة بيانات المجمعة مسبقا في هذه الدراسة اقترحنا تحسين لخوارزمية الجار الأقرب التي كانت أحسن الخوارزميات في نسبة التنبؤ ، لمحاولة رفع نسبة التنبؤ أكثر نظرا لخطر أمراض القلب على حياة الإنسان ،حيث أثبتنا فعالية الخوارزمية المحسنة بعد تدريبها على نفس البيانات التي دربنها بِها خوارزميات التعلم الآلي في نسبة التنبؤ.

الكلمات المفتاحية : أمراض القلب ، التعلم الآلي ، التنبؤ بأمراض القلب ، أقرب الجيران ، شجرة القرار ، الغابات العشوائية ، الانحدار اللوجستي

Abstract

Discovering a sicknesses before actually happens, counts as one of the most important factors at curing them. In the recent years, the medical field knew a huge expansion in the field of computer science as machine learning and deep learning where introduced, studied and heavily used to detect the different sicknesses. One of the most dangerous diseases to the human body are heart diseases, that's why it is a necessity to detect and cure them at early stages before they evolve to be more devastating. The goal of our project is to detect heart disease using supervised machine learning methods in the medical field. To do that we used different detecting algorithms, to find the algorithm with the most successful rate. After training them on pre-collected data set. In our study we suggested to improve the K nearest neighbour algorithm, which was the best detecting algorithm. We managed to prove the efficiency of our improved algorithm after using in on the same data set and getting better results than the other machine learning algorithms.

Key words : heart disease, machine learning, heart disease prediction, K nearest neighbors, decision tree, random forest, logistic regression.

Résumé

La détection prématurée des maladies cardiaques est l'un des facteurs les plus important qui font la différence dans le traitement des maladies, et c'est pourquoi, le monde assiste ces dernières années à différentes créations des méthodes d'apprentissage automatique et d'apprentissage profond pour prédire les maladies des personnes avant qu'elles ne surviennent. Parmi les causes de mortalité humaine, les maladies cardiaques font partie du podium, elles doivent donc être détectées et traiter avant qu'elles ne se développent.

Le but de notre travail est de détecter les maladies cardiaques, on utilisant l'apprentissage automatique supervisé. Pour ce fait, nous avons utilisé plusieurs algorithmes d'apprentissage automatique supervié, pour trouver le meilleur modèle pour un résultat maximal.

Dans notre étude, nous avons suggéré d'améliorer l'algorithme du K le plus proche voisin, qui était le meilleur algorithme de détection. Nous avons pu prouver l'efficacité de notre algorithme amélioré après avoir utilisé le même ensemble de données et obtenu de meilleurs résultats que les autres algorithmes d'apprentissage automatique supervisés.

Mots clés : les maladies cardiaques, apprentissage automatique, prédiction des maladies cardiaques, K plus proches voisins, arbre de décision, la foret aléatoire, la régression logistique

Table des matières

Table des matières **i**

Table des figures **v**

Liste des tableaux **vii**

Liste des abréviations **viii**

Introduction générale **1**

1 Généralités sur les maladies cardiaques **3**

1.1 Introduction 3

1.2 Le cœur 3

1.3 L'anatomie du cœur 4

1.4 Fonctionnement du cœur 5

1.5 Les maladies cardiovasculaires 5

1.5.1 Infarctus du myocarde (IDM) 6

1.5.2 Accident vasculaire cérébral (AVC) 6

1.5.3 Artériosclérose 6

1.5.4 Angine de poitrine – Angor 7

1.5.5 Insuffisance cardiaque 7

1.5.6 Arythmie cardiaque 7

1.5.7 Hypertension artérielle 7

1.6 L'examen du coeur 7

1.6.1 Méthode d'examen de maladie cardiovasculaire 8

1.7 Quelques conseils pour prévenir les maladies cardiovasculaires 10
1.8 Conclusion 11

2 L'apprentissage automatique 12
2.1 Introduction 12
2.2 Apprentissage automatique 12
2.3 Les types d'apprentissage automatique 13
2.3.1 Apprentissage Supervisé 13
2.3.2 Apprentissage semi-supervisé 14
2.3.3 Apprentissage par Renforcement 15
2.4 Quelques algorithmes de l'apprentissage automatique utilisés 16
2.4.1 Algorithme de plus proche voisine (KNN) : 16
2.4.2 Regression Logistique 19
2.4.3 Arbre de décision 19
2.4.4 Forets aléatoires (Random Forest) 21
2.5 Analyse du rendement et Evaluation des performances 24
2.6 Conclusion 25

3 Prédiction des maladies cardiaques par les techniques d'apprentissage automatique 26
3.1 Introduction 26
3.2 Problématique 26
3.3 Etat de l'art 27
3.4 Approche et solution proposée 29
3.4.1 Collecte des données 30
3.4.2 L'analyse exploratoire des données 31
3.4.3 Prétraitement de données 37
3.4.4 Entrainement de modèle 40
3.4.5 Optimisation de l'algorithme KNN 40
3.5 Conclusion 47

4 Résultats et évaluation 48
4.1 Introduction 48
4.2 Environnement de développement 48

4.2.1 La plateforme de développement Anaconda 48
4.2.2 La plateforme de développement Google Colab 49
4.2.3 Le langage de programmation python 49
4.2.4 Les bibliothèques utilisées 49
4.3 Résultats et analyse 50
4.3.1 Evaluation des performances de KNN, RF, DT, LR 50
4.3.2 Evaluation des performances de KNN régulier et KNN optimisé. . 52
4.4 Conclusion 55

Conclusion générale **56**

Bibliographie **58**

Table des figures

1.1 L'anatomie du cœur [1] 4
1.2 Circulation du sang dans le cœur[35] 5

2.1 Apprentissage automatique supervisé 13
2.2 Apprentissage par Renforcement 15
2.3 Un exemple simplifié d'un ensemble de formation 16
2.4 Une image montrant la deuxième étape du knn 17
2.5 une image montrant la troisième étape du knn 17
2.6 Structure de l'algorithme Arbre de décision (DT) 21
2.7 Diagrame forêts aléatoires 22

3.1 Architecture de notre projet 29
3.2 Rapport HTML de dataset 31
3.3 Visualisation de la classe cible 32
3.4 Visualisation relation âge /maladie cardiaque 33
3.5 Relation maladie cardiaque/sexe 33
3.6 Relation maladies cardiaque/Cp 34
3.7 Relation maladie cardiaque/thalach 34
3.8 Relation maladie cardiaque/la pression artérielle au repos 35
3.9 Matrice de corrélation 37
3.10 Les valeurs manquantes 38
3.11 Recherche de valeurs manquantes à l'aide de Heatmap 38
3.12 Le premier cas erroné du KNN 41
3.13 Le deuxième cas erroné prévu par le KNN 41

3.14 Le troisième cas erroné prévu par le KNN 42
3.15 Le quatrième cas erroné prévu par le KNN 42
3.16 exemple1 d'un cas erroné prévu par le KNN 44
3.17 exemple2 d'un cas erroné prévu par le KNN 45
3.18 exemple3 d'un cas erroné prévu par le KNN 45

4.1 La comparaison de la précision des différents algorithmes. 51
4.2 Comparaison de l'exactitude pour différents modèles. 55

Liste des tableaux

2.1 Illustration d'une matrice de confusion 24

3.1 Comparaison des travaux étudiés. 28
3.2 Les attributs de dataset 30

4.1 la matrice de confusion des différents modèles. 50
4.2 les métriques de performance de chaque algorithme. 51
4.3 Tableau d'analyse pour déterminer la meilleure valeur K pour KNN et KNN optimisé 52
4.4 Tableau d'analyse pour déterminer la meilleure valeur K pour KNN régulier et KNN optimisé. 53
4.5 matrice de confusion des algorithme KNN, DT, LR, RF et KNN-optimisé. 54
4.6 comparison de l'exactitude. 54

Liste des abréviations

ML	Machine Learning
LR	Logistic Regression
DT	Decision Tree
RL	Régréssion Logistique
RF	Random Forest
TP	Ture Positive
FP	False Negative
TN	True Negative
FN	False Negative
IDM	Infarctus du myocarde
KNN	K Nearest Neighbour
AVC	Accident vasculaire cérébral
ECG	électrocardiogramme
ASA	administration d'aspirine
LDL	mauvais colesterole
IA	Intelligence artificielle
OOB	out of bag
UCI	Heart Disease Dataset
MLP	perceptron multicouche

Introduction générale

Les soins de santé sont la perpétuation de l'état de bonne santé d'une personne par la prévention, le diagnostic et le traitement des maladies. Selon un vieil adage, « La santé humaine est une richesse », un corps sain est la chose la plus précieuse qu'une personne puisse posséder.

Le cœur est l'un des organes centraux et les plus vitaux du corps humain, il joue un rôle crucial dans le pompage du sang dans le corps humain qui est aussi essentiel que l'oxygène pour le corps humain, il est donc toujours nécessaire de le protéger. Les maladies cardiaques sont aujourd'hui l'une des causes de mortalité les plus importantes dans le monde, la plupart des patients sont décédés parce que leurs maladies sont reconnues au dernier stade en raison du manque de précision des instruments utilisés, La prédiction des maladies cardiovasculaires est un défi critique dans le domaine de l'analyse des données cliniques. L'apprentissage automatique s'est avéré efficace pour aider à prendre des décisions et des prédictions à partir de la grande quantité de données produites par le secteur de la santé.

L'apprentissage automatique est une technique d'analyse de données qui automatise la création de modèles pour aider à résoudre plusieurs problèmes dans notre vie quotidienne. 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. Quelques exemples : détection de visage reconnaissance vocale, prédiction structurée, détection d'anomalies... etc.

Les algorithmes d'apprentissage automatique peuvent être classés en quatre types (i) apprentissage non supervisé ; (ii) apprentissage supervisé ; (iii) apprentissage semi-

supervisé ; et (iv) apprentissage par renforcement. Fondamentalement aucune des données n'est étiquetée dans le type (i), les échantillons d'entrainement/test sont étiquetés dans le type (ii) et il y a beaucoup de données non étiquetées et peu de données étiquetées dans le type (iii).

L'objectif de ce projet est de vérifier si le patient est susceptible être diagnostiqué avec des maladies cardiaques en fonction de leurs attributs médicaux tels que le sexe, l'âge, les douleurs thoraciques, le taux de sucre à jeun, etc. Un ensemble de données est sélectionné dans le référentiel UCI avec les antécédents médicaux et les attributs du patient. En utilisant cet ensemble de données, nous prédisons si le patient peut avoir une maladie cardiaque ou non. On a utilisé dans ce projet quatre algorithmes qui sont arbre de décision (DT), régression logistique (RL), k plus proche voisin (KNN), la foret aléatoire (RF). On a calculé le taux d'exactitude de chaqu'un de ces quatre algorithmes et on a constaté que le plus efficace de ces algorithmes est le KNN qui nous donne un taux d'exactitude de 91,42 %, puis on a optimisé cet algorithmes pour atteindre un taux d'exactitude qui vaut 95,71%

Pour atteindre notre objectif, notre mémoire est organisé comme suit :

Le premier chapitre, présente des généralités sur les maladies cardiaques.

Dans le deuxième chapitre, nous avons présenté l'apprentissage automatique en définissant ses différents types ainsi ses algorithmes qui nous ont permis de modéliser un modèle d'apprentissage automatique.

Le troisième chapitre représente notre proposition (solution) qui a été expliquée et détaillée en passant par des étapes essentielles afin d'arriver à son implémentation finale. Le dernier chapitre présente la discussion des résultats obtenus et l'implémentation de notre proposition.

Nous terminerons ce mémoire par une conclusion générale qui résume le travail effectué.

Chapitre 1

Généralités sur les maladies cardiaques

1.1 Introduction

Le nombre de personnes souffrant des maladies cardiaques est en augmentation. Un diagnostique précis à un stade précoce suivi d'un traitement ultérieur approprié peut permettre de sauver un nombre de vies considérable. Pour diagnostiquer et prédire les maladies cardiaques, nous devons avoir une bonne connaissance du cœur et son fonctionnement. Dans ce qui suit nous allons décrire l'anatomie du cœur et son fonctionnement, les maladies cardiaques et les facteurs de risque.

1.2 Le cœur

Le cœur est un muscle d'une taille d'environ d'un poing, il est protégé par la cage thoracique notamment par les cotes et le sternum. Sa fonction consiste à faire circuler le sang dans le corps. Il bat en moyenne 100 000 fois par jour et à chaque battement, le cœur pompe le sang et le fait circuler dans le réseau d'artères et de veines. Le sang achemine l'oxygène et les nutriments essentiels vers chaque cellule du corps. Les artères distribuent le sang riche en oxygène provenant du cœur dans le corps et les veines ramènent le sang pauvre en oxygène au cœur et aux poumons [1].

1.3 L'anatomie du cœur

Le cœur se présente sous le format d'une pyramide triangulaire, il est ferme et rouge. Il pèse chez l'homme 300 g et chez la femme 270 g. Il est composé de quatre parties, appelés les cavités. On appelle les deux cavités supérieurs les oreillettes, et les cavités inférieures les ventricules. Une paroi musculaire appelée septum sépare les côtés droit et gauche du cœur[2].

Le côté droit du cœur, où se trouvent l'oreillette et le ventricule droits, reçoit le sang appauvri en oxygène provenant du reste du corps. Le côté gauche, où se trouvent l'oreillette et le ventricule gauches, reçoit le sang fraîchement oxygéné par les poumons[3] Les quatre cavités communiquent entre elles au moyen de valvules, ou valves, qui s'ouvrent et se referment chaque battement du cœur. Ainsi, il y a quatre valvules cardiaques : la valvule aortique, la valvule tricuspide et la valvule mitrale. Ces valvules antiretours imposent un sens unique à la circulation du sang, qui passe d'une cavité du cœur à l'autre, et ainsi de suite, pour ensuite être expulsé vers le reste du corps. Les « battements » du cœur qu'on arrive à entendre avec le stéthoscope sont en fait l'ouverture et la fermeture des valvules cardiaques pour laisser passer le sang [3].

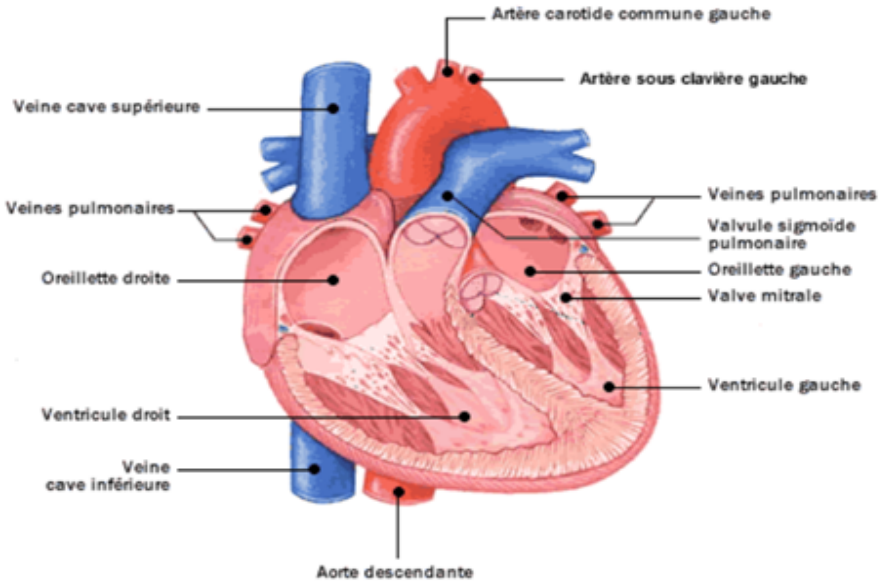


FIGURE 1.1 – L'anatomie du cœur [1]

1.4 Fonctionnement du cœur

1. Le côté droit du cœur renvoi le sang pauvre en oxygène aux poumons pour éliminer le dioxyde de carbone et ré oxygéner 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. [4]

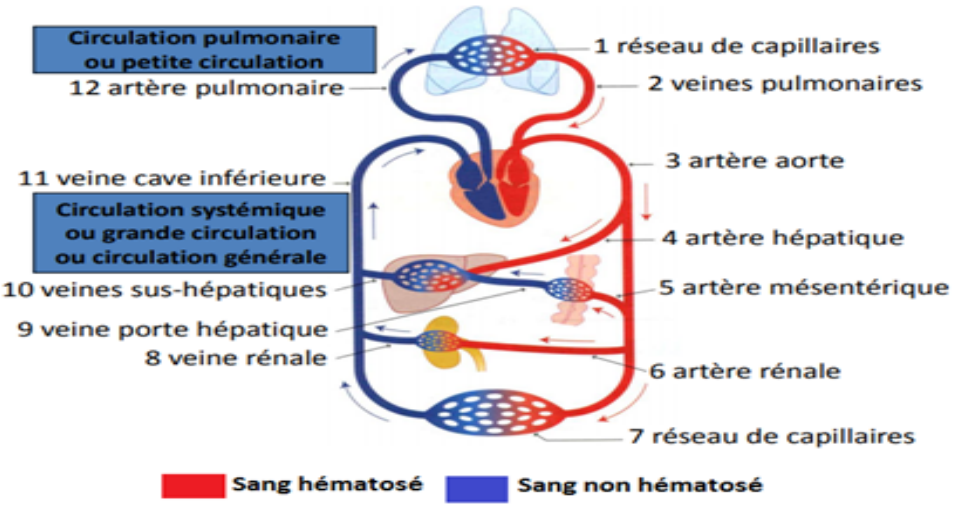


FIGURE 1.2 – Circulation du sang dans le cœur[35]

1.5 Les maladies cardiovasculaires

Une maladie cardiovasculaire est une pathologie qui touche le cœur et les vaisseaux sanguins. Les maladies cardiovasculaires sont la première cause de mortalité dans le monde et en Algérie avec un taux de 34% par an selon les chiffres de l'Institut national de la

santé publique (INSP)". [5] Dix facteurs de risques déterminés par le mode de vie (et donc potentiellement modifiables) sont associés à la survenue de 90% des AVC :

- L'antécédent d'hypertension artérielle, qui contribue à 40% au risque d'AVC (risque multiplié par 2, et par 5 chez les moins de 55 ans) ;
- Le tabagisme, qui triple le risque d'AVC ;
- L'obésité abdominale, évaluée par le rapport du tour de taille/tour de hanche, qui contribue à hauteur de 36% à l'AVC ;
- Une alimentation non équilibrée contribue à hauteur de 33% au risque d'AVC ;
- Le manque d'activité physique ;
- La consommation d'alcool ;
- La fibrillation atriale, qui est le premier facteur de risque d'origine cardiaque, avec un risque multiplié par 4 ;
- Les facteurs psychosociaux (stress, dépression, isolement social...) ;
- Un diabète, pour l'AVC ischémique ;
- Une concentration trop élevée d'un ou plusieurs lipides présents dans le sang (cholestérol, triglycérides...).

1.5.1 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 [6][7].

1.5.2 Accident vasculaire cérébral (AVC)

L'accident vasculaire cérébral (AVC) est l'arrêt brutal de la circulation sanguine dans une ou plusieurs parties du cerveau. La rapidité de prise en charge est essentielle. Le symptôme le plus courant d'un AVC est une sensation de faiblesse soudaine au niveau de la face, du bras ou de la jambe, le plus souvent sur un seul côté du corps [6][7].

1.5.3 Artériosclérose

L'artériosclérose est un épaississement et un durcissement de la paroi des artères [6][7].

1.5.4 Angine de poitrine – Angor

L'angine de poitrine décrit une douleur violente localisée dans la poitrine [6][7].

1.5.5 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, surpoids, diabète...) en modifiant durablement ses habitudes de vie [6][7].

1.5.6 Arythmie cardiaque

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

1.5.7 Hypertension artérielle

L'hypertension artérielle est une augmentation de la pression du sang dans les artères [6][7].

1.6 L'examen du coeur

L'examen physique est l'un des composés vitaux de la série d'actions visant à déterminer la présence ou l'absence de la maladie, sa gravité et l'étendue du handicap du patient. Afin de déterminer le traitement approprié. Cette procédure comprend l'audition de la plainte du patient et l'histoire de sa maladie, un examen physique, ainsi que des examens spéciaux (par exemple ECG), une imagerie et un Doppler échographique (Echo - doppler). Comme d'autres systèmes, l'examen physique du cœur et des vaisseaux sanguins dépend principalement des différents sens du médecin examinant le patient, notamment : la vision, le toucher, la percussion (tapotement) et l'écoute, qui nécessitent une grande compétence. Et ce malgré la légère baisse de l'importance de ce choix au cours des dernières années, par rapport aux tests spécialisés qui ont été développés. Il convient de noter que ce test est disponible et disponible en permanence chez le médecin, partout et à tout moment, et n'implique la présence d'aucun appareil pouvant être nécessaire pour effectuer le test (à l'exception du stéthoscope et du tensiomètre) et les résultats de ce test sont immédiats

(critiques dans les cas critiques). Des données plus complètes qui peuvent être déduites par ce test que des données qui peuvent être obtenues par des tests spéciaux individuels. Par exemple, le diagnostic de sténose valvulaire se fait souvent par examen physique, tandis qu'un ECG ne peut que soulever des doutes sur la possibilité d'une sténose valvulaire [6][7].

1.6.1 Méthode d'examen de maladie cardiovasculaire

- Inspection En regardant la peau et les muqueuses, il est possible d'estimer les oxydations sanguines artérielles insuffisantes (qui peuvent résulter de malformations cardiaques congénitales, de maladies pulmonaires et d'insuffisance cardiaque aiguë). Cela apparaît sous la forme de feuilles, la couleur de la peau et des muqueuses, tandis que l'âge d'apparition de la maladie peut indiquer sa cause [6][7].
- Diagnostique d'anémie La pâleur de la peau indique la présence d'anémie, un manque d'hémoglobine porteuse d'oxygène dans le sang - qui peut être associé à une maladie cardiaque, telle qu'une inflammation de la coque interne du cœur due à une infection bactérienne. Dans les cas où l'essoufflement devient plus léger en position assise par rapport à la position couchée, il soulève des soupçons que le patient peut avoir une insuffisance cardiaque du côté gauche. En outre, le schéma respiratoire peut nous fournir des informations sur le type de maladie cardiaque [6][7].
- **La peau :** L'apparence, la forme et la position des taches rouges sur la peau peuvent soulever des soupçons que le patient peut avoir diverses maladies cardiaques aiguës, telles que les maladies cardiaques rhumatismales aiguës ou l'inflammation bactérienne de la muqueuse interne du cœur.
- **Les anomalies squelettiques :** Les anomalies squelettiques, telles que l'angiome d'araignée, ou l'apparition d'un sixième doigt dans la main, peuvent être associées à un certain nombre d'anomalies cardiaques différentes [6][7].
- **L'œdème :** L'œdème (œdème) est un gonflement causé par une accumulation de liquide-peut être un signe d'insuffisance cardiaque droite ou de maladies qui rendent plus difficile le retour du sang dans le ventricule droit (péricardite constrictive) et la sténose tricuspide (sténose tricuspide)[6][7].
- **Palpitations veineuses :** Prévisualiser les palpitations veineuses dans le cou

peut nous fournir beaucoup d'informations : la vitesse et la régularité des impulsions, la hauteur de la colonne sanguine dans la veine en position assise, comment elle se déplace en expirant et en inspirant [6][7].

— **Toucher par la paume de la main du médecin :** Même aujourd'hui, ce processus est considéré comme une partie importante du test. La palpation de la zone de la cage thoracique aide le médecin à ressentir les vibrations causées par une sténose valvulaire aiguë ou la sensation de grattage pouvant résulter d'une péricardite [6][7].

— **Citrouille(Percussion) :** Utilisé même il y a 30 ans, c'est une tentative de comprendre grossièrement les limites du cœur. La percussion sur la cage thoracique dans la région du cœur a donné une réaction de déglutition sonore (profonde), par opposition à l'écho qui survient lorsque la cage thoracique est une percussion dans la région des poumons. De nos jours, les citrouilles ne sont utilisées que pour évaluer les poumons et détecter s'il y a du liquide accumulé dans la plèvre qui entoure les poumons. La découverte de ce liquide soulève des soupçons que le patient a une maladie cardiaque qui rend plus difficile le retour du sang dans le ventricule droit, ou même d'autres développements cardiaques ou pulmonaires [6][7].

— **Écoute (Auscultation) des sons et des bruits du cœur :** Il constitue une pierre angulaire de ce test. Dans un tel cas, la compétence du médecin concerne sa capacité à écouter séparément chacun des composés suivants : sons cardiaques (en général, il y a deux sons cardiaques par cycle, mais il est possible d'entendre 4 sons), bruit systolique - pendant la contraction cardiaque, diastolique - pendant la diastole cardiaque (expansion), son de friction (Frottement) et autres sons. Les sons du cœur causés par l'ouverture et la fermeture des valves et la fluctuation des parois du cœur et des principaux vaisseaux sanguins, ainsi que la force, le rythme et la forme du bruit, aident le médecin à déduire la présence d'une maladie valvulaire spéciale (sténose ou insuffisance valvulaire) le distinguer, évaluer sa gravité et, dans certains cas, Le bruit de friction et d'autres sons peuvent indiquer une maladie péricardique. L'écoute de la surface de l'abdomen ou du dos peut aider à diagnostiquer d'autres blessures vasculaires.

— **Les résultats de l'examen physique** doivent être documentés dans un rapport médical, en mettant l'accent sur les changements survenus depuis le dernier examen physique.[6][7]

1.7 Quelques conseils pour prévenir les maladies cardiovasculaires

Les maladies cardiaques peuvent être réduites en prenant des médicaments préventifs, une alimentation saine, des sports, en évitant le tabagisme et l'alcool et en gérant le stress et le stress, qui aident tous à prévenir les maladies cardiaques [8].

Médicaments préventifs : La recherche clinique chez l'homme a montré que l'administration d'aspirine-ASA - anticoagulant aux patients atteints d'infarctus du myocarde (prévention secondaire) et aux patients dont l'athérosclérose n'a pas encore causé de lésions tissulaires (prévention primaire) réduit le risque d'infarctus du myocarde d'environ 35%, réduit la probabilité d'arrêt cardiaque d'environ 23% et réduit la probabilité de décès En revanche, l'administration de médicaments anticoagulants, tels que l'aspirine, implique un risque accru de saignement dans l'estomac, mais seulement d'un faible pourcentage [8].

Une alimentation saine et équilibrée : Le mauvais cholestérol (LDL) est le facteur numéro un que nous devons maintenir dans la plage normale, et nous devons également éviter les triglycérides élevés,Comme les médecins conseillent de manger des légumes, des fruits, du poisson, des produits alimentaires à base de soja, des produits laitiers , des produits alimentaires faibles en gras riches en fibres, et les médecins avertissent de la consommation fréquente de viande, de produits de boulangerie, de produits laitiers et de restauration rapide riches en graisses saturées, vous devez également limiter la consommation de glucides présents dans le pain blanc, les pommes de terre, le chocolat, la carotte, la pastèque, le riz blanc, les bonbons, les aliments frits et les aliments riches en sel [8].

Exercice physique : Les médecins recommandent de choisir un certain type de sport physique et d'y adhérer plusieurs fois par semaine, ils conseillent également de faire du sport une partie de notre vie quotidienne, lors du trajet, il est recommandé de marcher

au lieu de monter dans les transports et de monter les escaliers au lieu d'ascenseur, et la chose la plus importante pour [8].

Arrêter de fumer : La recherche montre que le tabagisme augmente le risque de maladie cardiaque et pulmonaire, de divers cancers, d'affaissement de la peau et d'accélération du vieillissement, d'accident vasculaire cérébral et de démence précoce [8].

Alcool : L'alcool est un sucre qui peut pénétrer directement dans le sang, fait augmenter le taux de triglycérides dans le sang et affecte négativement l'hypertension. Il est préférable de s'abstenir de consommer de l'alcool et de recourir à des moyens plus sains pour augmenter le taux de bon cholestérol [8].

Pression et tension : Il n'y a pas de médicament pour se débarrasser du sentiment de stress et de stress une fois pour toutes, il existe des techniques de relaxation telles que le yoga et la méditation, en plus d'un bon sommeil et de repos, créant un environnement favorable qui comprend des amis et des proches pour partager avec eux nos problèmes lorsque nous sommes bouleversés.[8]

1.8 Conclusion

Dans ce premier chapitre, nous avons d'abord décrit l'anatomie du cœur et son fonctionnement puis nous avons donné un aperçu sur les maladies cardiaques et ses facteurs de risque et à la fin nous avons donné quelques conseils pour la prévention des maladies cardiovasculaires. Dans le second chapitre, nous présenterons des approches différentes d'aide au diagnostic préventif en utilisant des algorithmes d'apprentissage automatique dans la prédiction des maladies cardiaques.

Chapitre 2

L'apprentissage automatique

2.1 Introduction

L'apprentissage automatique permet à la machine d'évoluer à travers un processus d'apprentissage pour effectuer des tâches complexes qui ne sont pas explicitement programmées via l'apprentissage des données, La technologie de l'IA est en plein essor et se répand largement maintenant avec l'avènement du big data ,dans ce chapitre, nous identifierons les principaux types d'apprentissage automatique et les algorithmes utilisés, et nous fournirons donc des recherches sur l'application des algorithmes d'apprentissage automatique et pour augmenter le taux de prédiction au plus haut niveau possible des maladies cardiaques, afin de réduire le risque de complications sur la santé du patient, et rapprocher l'hôpital du patient

2.2 Apprentissage automatique

L'apprentissage automatique est une application de l'intelligence artificielle qui permet aux systèmes d'apprendre avec expérience et de s'améliorer sans être explicitement programmés pour cela. L'apprentissage automatique se concentre sur le développement de programmes informatiques qui peuvent accéder aux données et les utiliser pour apprendre par eux-mêmes. Pour mieux comprendre le fonctionnement de l'apprentissage automatique, il faut comprendre comment former un problème d'apprentissage. Tom M. Mitchell le définit bien dans son livre intitulé «Machine Learning» comme tel :« un programme informatique apprend de l'expérience E par rapport à une classe de tâches T et

à des objectifs de performance P lorsque ses performances sur des tâches en T, mesurées par P, s'améliorent avec l'expérience e »[11].

2.3 Les types d'apprentissage automatique

L'apprentissage automatique implique différents types d'apprentissage, en fonction de la structure du problème. Les plus courants dans les domaines d'application dominants sont l'apprentissage supervisé et non supervisé, ainsi que l'apprentissage semi-supervisé et par Renforcement.

2.3.1 Apprentissage Supervisé

Les méthodes d'apprentissage supervisées exigent que la valeur des variables de sortie pour chaque test d'apprentissage soit connue. En conséquence, chaque test d'entraînement est présenté sous la forme d'une paire de valeurs d'entrée et de sortie. L'algorithme forme ensuite un modèle qui prédit la valeur des variables de sortie à partir des variables d'entrée en utilisant les fonctions définies dans le processus. Si la variable de sortie est évaluée en continu, le modèle de prédiction est appelé fonction de régression. Par exemple, prédire la température de l'air à une période particulière de l'année est un problème de régression. Si la variable de sortie est un ensemble discret de valeurs, le modèle de prédiction est appelé classificateur. Un problème de classification typique est le diagnostic médical automatisé, pour lequel les données d'un patient doivent être classées comme une maladie ou non[9].

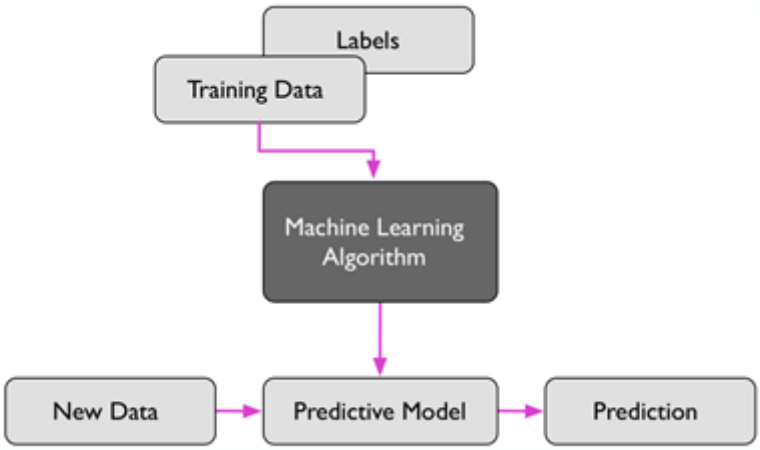


FIGURE 2.1 – Apprentissage automatique supervisé

Différence entre classification et régression

Il y a une grande différence entre les problèmes de classification et de régression. La classification est utilisée pour prédire une étiquette, alors que la régression est utilisée pour prédire une quantité. Avant d'introduire ces deux concepts, il est utile de comprendre le concept d'approximation de fonction.

— **Approximation de la fonction :** : La modélisation prédictive désigne un ensemble de méthodes permettant l'analyse et l'interprétation de données définies pour prédire les données futures. La modélisation prédictive peut être décrite comme un problème mathématique d'approche d'une fonction de mappage f entre les variables prédictives d'entrée X et la variable prédictive Y. C'est ce qu'on appelle la fonction d'approximation du problème. en général, toutes les tâches d'approximation peuvent être divisées en tâches de classification et en tâches de régression[14].

— **Classification :** : Les algorithmes de classification sont utilisés lorsque la variable à prédire est discrète. Une classification peut avoir des variables d'entrée réelles ou discrètes. Il est courant que les modèles de classification prédisent des valeurs continues sous forme de probabilités d'appartenance à chaque classe de production. Une probabilité prédite peut être convertie en une valeur de classe en sélectionnant le nom de la classe avec la probabilité la plus élevée [14].

— **Régression** Les algorithmes de régression sont utilisés lorsque la variable à prédire est continue. Comme pour prédire le prix d'une voiture. Un problème de régression nécessite l'anticipation d'un ensemble. Il peut avoir des variables d'entrée à valeur réelle ou discrètes. Notez que certains algorithmes qui ont le mot" régression " dans leur nom, tels que la régression linéaire et la régression logistique, peuvent prêter à confusion car la régression linéaire est un algorithme de régression, tandis que la régression logistique est un algorithme de classification. Certains algorithmes peuvent être utilisés à la fois pour la classification et la régression avec de petites modifications, telles que les arbres de décision et les réseaux de neurones artificiels [14].

2.3.2 Apprentissage semi-supervisé

L'apprentissage semi-supervisé se situe au milieu des deux méthodes mentionnées ci-dessus et peut être plus avantageux, car les données non marquées sont plus accessibles

que les données de haute qualité marquées. Cette famille de méthodes d'apprentissage fonctionne avec un petit ensemble de données d'entraînement marquées(supervisé)et un plus grand ensemble de données non marquées(non supervisé). Lors de la formation d'un modèle de prédiction, ces algorithmes peuvent utiliser à la fois des valeurs de sortie surveillée et la distribution des données en données non marquées. Cependant, ces algorithmes font des hypothèses supplémentaires pour tirer parti de données non marquées qui peuvent ou non convenir au problème en question.[9]

2.3.3 Apprentissage par Renforcement

La théorie du contrôle optimal a permis la naissance d'un type d'apprentissage qui est l'apprentissage par renforcement, une approche légèrement différente des méthodes ci-dessus les théories logiques et les modèles de probabilité à partir d'exemples. L'apprentissage par renforcement, il s'agit de savoir comment les agents peuvent apprendre quoi faire sans exemples étiquetés. C'est une méthode d'apprentissage qui interagit avec l'environnement en générant et en découvrant des actions d'échec ou de récompenses environnementales. La recherche par essais et erreurs, tout en faisant face au dilemme exploitation et exploration, et le défi de la modélisation des récompenses retardées sont quelques-unes des principales caractéristiques de l'apprentissage par renforcement. Cette méthode permet aux machines et agents flexibles de déterminer automatiquement le comportement idéal dans un contexte donné afin de maximiser les performances de l'agent.. Un simple retour de récompense est nécessaire pour que l'agent puisse savoir quelle action est la meilleure. Le fait qu'il n'ait pas besoin d'être marqué les exemples le rendent différent et précieux[10][11].

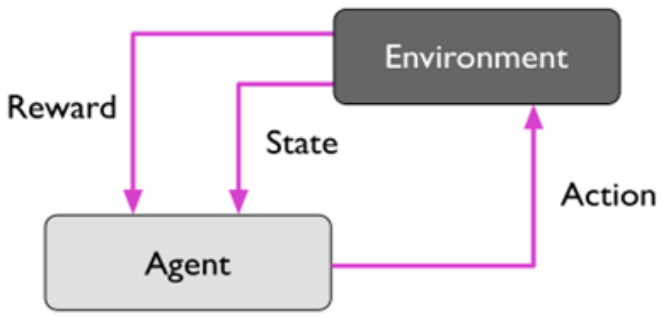


FIGURE 2.2 – Apprentissage par Renforcement

2.4 Quelques algorithmes de l'apprentissage automatique utilisés

2.4.1 Algorithme de plus proche voisine (KNN) :

L'algorithme des k plus proches voisins s'écrit en abrégé k-NN ou KNN (de l'anglais k-nearest neighbors), C'est un type d'algorithme d'apprentissage automatique supervisé qui traite des problèmes de classification et de prédiction. Il s'agit d'un algorithme d'apprentissage paresseux, car il n'a pas d'étape d'apprentissage spéciale. Au lieu de cela, il utilise toutes les données pour la formation lors de la classification ou de la prédiction d'un nouveau point de données, Puisque le problème que nous avons est un problème de classification en deux cellules, c'est-à-dire si le patient a une maladie cardiaque ou non, nous expliquerons en profondeur l'algorithme du voisin le plus proche de la classification

Exemple avec explication :

Disons que l'ensemble des barres spéciales du carrelage sont les carrés rouges et les triangles verts, et le cas que nous nous attendons à classer est le cercle noir avec un point d'interrogation. Comme il est montré dans la figure suivante.

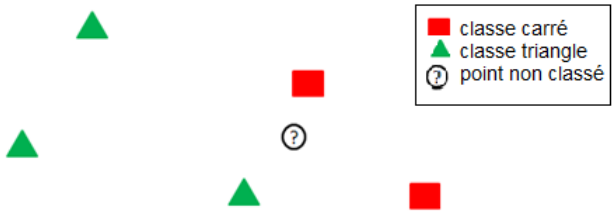


FIGURE 2.3 – Un exemple simplifié d'un ensemble de formation

La première opération de l'algorithme consiste à calculer les distances entre le point testé et tous les points d'entraînement existants en utilisant plusieurs fonctions existantes et les distances sont organisées du plus petit au plus grand comme indiqué.

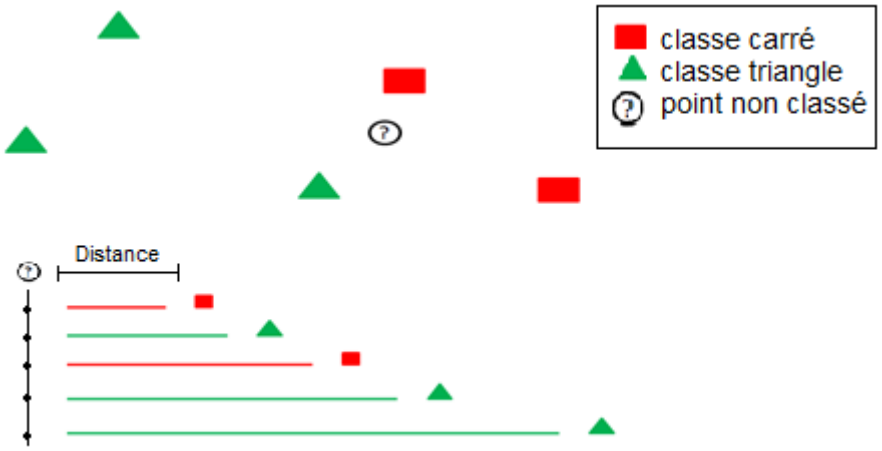


FIGURE 2.4 – Une image montrant la deuxième étape du knn

Les distances entre le point d'essai et les points d'entraînement sont disposées. Les données k sont conservées pour l'ensemble de données le plus proche de la valeur testée et disons K=3.

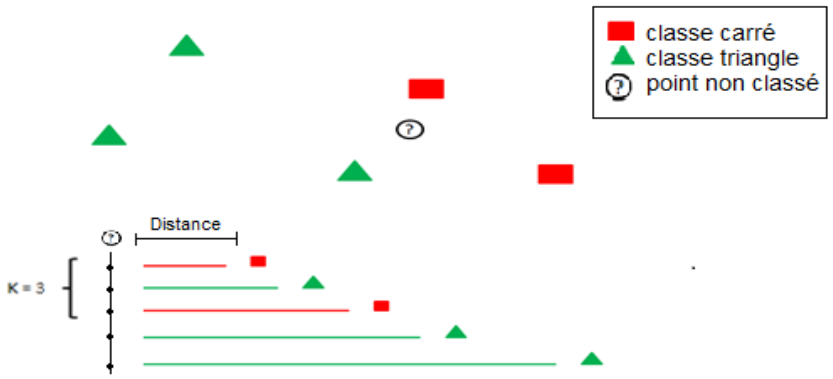


FIGURE 2.5 – une image montrant la troisième étape du knn

La catégorie la plus courante est attribuée aux données K les plus proches. Dans ce cas, le nombre de carrés rouges K prédomine, de sorte que l'algorithme du voisin le plus proche détermine que le cas testé est un carré rouge.

Fonctions mises en évidence pour le calcul des distances

. Les voisins les plus proches sont déterminés par les distances les plus faibles mesurées de la manière suivante :

1. *Distance euclidienne* : la distance euclidienne est calculée comme la racine carrée de la somme des carrés des différences entre le nouveau point et le point existant

pour tous les attributs d'entrée. C'est le plus largement utilisé.

$$\sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

2. *Distance de Manhattan* : La méthode de calcul de la distance entre le point testé et un point de l'ensemble de test est la somme des différences absolues des coordonnées cartésiennes.

$$\sum_{i=1}^{k} x_i - y_i$$

3. *Distance de Hamming* : Est l'idée de Richard hemming, et utilisé dans ordinateur. Leur but est de déterminer la différence entre deux séquences de symboles. C'est distance au sens mathématique du terme. Avec deux séquences de symboles de même longueur, il relie Le nombre de positions où les séquences diffèrent.

Il existe plusieurs autres façons de calculer la distance, telles que la distance de Jaccard en Cosine. Le plus important est celui que nous avons mentionné dans le début.

Algorithme de construction de KNN

- **Étape 1** : Sélectionnez le nombre K de voisins
- **Étape 2** : Calculez la distance
- **Étape 3** : Prenez les K voisins les plus proches selon la distance calculée.
- **Étape 4** : Parmi ces K voisins, comptez le nombre de points appartenant à chaque catégorie.
- **Étape 5** : Attribuez le nouveau point à la catégorie la plus présente parmi ces K voisins[33].

Avantage de KNN

- L'algorithme est simple et facile à mettre en œuvre.
- Aucune hypothèse sur les données (linéaires, affines,...).
- L'algorithme est polyvalent. Il peut être utilisé pour la classification, la régression.

Inconvénients de KNN

- L'algorithme devient beaucoup plus lent à mesure que le nombre d'exemples d'apprentissage augmente.

— Le choix de la méthode de calcul de la distance ainsi que le nombre de voisins K peut ne pas être évident [34].

— Stockage de données.

— Ses défauts peuvent être observés sur le graphique.

2.4.2 Regression Logistique

La régression logistique est très proche de la régression linéaire. La régression linéaire permet de caractériser les relations entre une variable à expliquer (Y) variables quantitatives et explicatives (X1,X2, X3,... Xn) utilisation du modèle selon la formule suivante[12][13][19] :

$$Y = \beta_0 + \beta_1 X1 + \beta X2 + \ldots \beta_n Xn + \epsilon$$

Si la variable dépendante Y est binaire, alors le modèle de régression linéaire n'est pas le plus approprié . Il est donc nécessaire d'utiliser un modèle plus adapté qui permette de relier les variables explicatives à la variable qualitative Y. L'astuce de la régression logistique n'est pas de modéliser la variable qualitative Y, mais la probabilité que la variable dépendante prenne les valeurs 1 ou 0. [12][13][19]

Le modèle logistique permet une expression non linéaire qui varie de 0 à 1. La régression logistique de est basée sur l'hypothèse suivante :

$Evidence = \frac{lnp}{(1-p)} = LOGIT = \beta_0 + \beta_1 X1 + \beta X2 + \ldots \beta_n Xn + \epsilon$

où p est la distribution conditionnelle de X qui connaît la valeur de Y. par transformation logarithmique, nous obtenons

$$P(Y) = \frac{1}{1 + e^{(\beta_0 + \beta_1 * X1 + \beta_2 + \beta_2 * X1 + \ldots\ldots \beta_n * X_n)}}$$

P(y) est la probabilité que cela se produise. Si la valeur prévue est supérieure à 0,5, l'événement est susceptible de se produire alors que cette valeur est inférieure à 0,5[19].

2.4.3 Arbre de décision

Les arbres de décision 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 (séquences de décision individuelles par segment).groupe) pour prédire un résultat ou une classe.

En théorie des graphes, un arbre est un graphe non orienté, acyclique et connexe. Ensemble le nœud est divisé en trois catégories :

— **Nœud racine** : Ce nœud est utilisé pour accéder à l'arborescence.

— **Nœuds internes** : nœuds qui ont des descendants.

— **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 testées dans les nœuds de l'arbre. Les tests sont effectués dans les nœuds internes et les décisions sont prises dans les nœuds feuilles[15].

Apprentissage avec les arbres de décision

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

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

— Variable numérique : Test par intervalles de valeurs.

Cela permet aux feuilles de l'arbre de spécifier les classes et d'encoder la règle de décision. Lorsque l'arbre est construit, la classification d'un nouvel individu s'effectue par une descente dans l'arbre, de la racine à l'une des feuilles. À chaque niveau de la descente, un nœud intermédiaire est passé où une variable est testée pour déterminer le chemin à choisir pour continuer la descente[15].

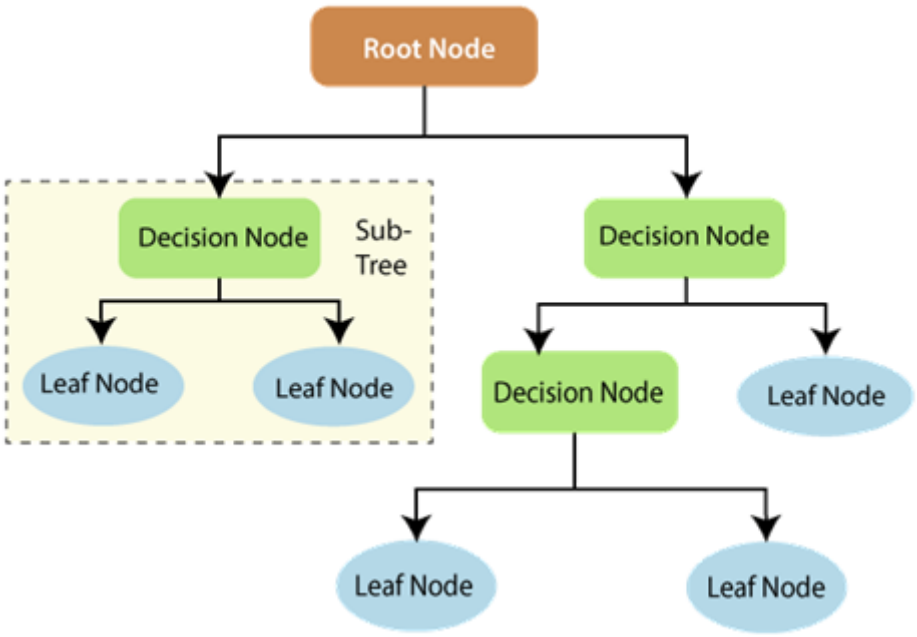


FIGURE 2.6 – Structure de l'algorithme Arbre de décision (DT)

2.4.4 Forets aléatoires (Random Forest)

Les algorithmes de forêt aléatoire sont des algorithmes d'apprentissage supervisés introduits par Breiman. L'algorithme de structure aléatoire est connu pour la précision de sa prédiction et sa capacité à exécuter des bases de données avec un nombre réduit d'observations et un grand nombre de variables explicatives. Cette technique est utilisée pour améliorer le pouvoir prédictif dans un arbre de décision unique. L'algorithme crée des arbres indépendamment, et à chaque fois un rééchantillonnage aléatoire de la base d'apprentissage est utilisé ,appelé Boosting [15].

Bagging (ou bootstrap aggregating)

Soit Y la variable à expliquer, les variables explicatives et $\Omega le modèle apprissur un échantillon z = \{(x_i, y_i) \dots (x_n, y_n)\}$. Le bagging consiste en les étapes suivantes :

1. On considère B échantillons bootstrap $z_i, ..., z_n$d'individus. Ces échantillons de z sont tirés au hasard par tirage au sort avec remise.

2. Sur chacun des échantillons, nous apprenons un modèle $\Omega(z_i)$.

3. On prédit Y en agrégeant les différentes décisions sur chacun des z_i par

$$Y = \Omega(x) = \frac{1}{B}\sum_{i=1}^{n} \Omega_{zi}(x)$$

Notez que ce type d'agrégation de résultats n'affecte que la régression.. Pour une classification nous prenons la décision majoritaire.

Algorithme forêts aléatoires :

- Nous tirons au hasard dans la base d'apprentissage B échantillons avec remise (chaque échantillon ayant n points).
- Pour chaque échantillon i, nous construirons construit un arbre CART selon un algorithme légèrement modifié : a chaque fois qu'un nœud doit être coupé (étape "split") nous tirons au hasard une partie des attributs (q parmi les p attributs) et nous choisissons le meilleur découpage dans ce sous-ensemble.
- Régression : agrégation par la moyenne $\Omega(x) = \frac{1}{B}\sum_{i=1}^{B} \Omega_i(x)$
- Classement : agrégation par vote $\Omega(x) = \text{Vote majoritaire}\,(\Omega_1(x), \ldots, \Omega_B(x))$
- Chaque arbre est petit donc moins performant, mais l'agrégation compense pour ce manquement (chaque attribut se retrouve typiquement dans plusieurs arbres) [14][15].

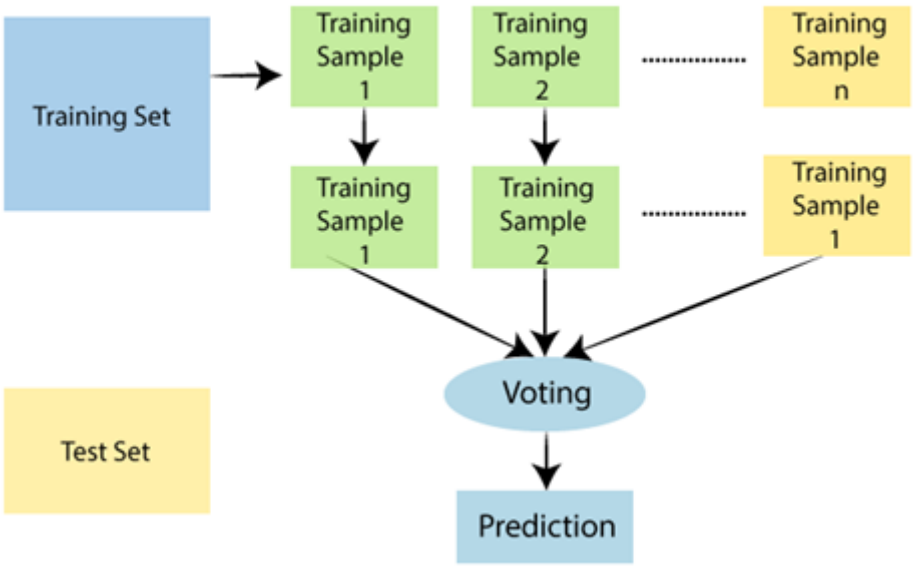


FIGURE 2.7 – Diagrame forêts aléatoires

L'importance des attributs :

Les attributs peuvent être évalués pour déterminer leur impact sur la structure arborescente (mesure de Gini) ou la résistance à l'erreur de captures ou au bruit lors de la classification (erreur OOB) : Gini : Le changement dans l'impureté (ou la réception d'informations) dans cumulé sur tous les arbres de la forêt. Erreurs OOB : Tous les échantillons OOB sont évalués à l'aide d'arbres et d'erreurs de mesure. Ensuite, nous changeons aléatoirement les valeurs de chaque attribut J et mesurons à nouveau la fréquence d'erreur. La valeur finale est la dégradation moyenne (changement d'erreur) de tous les arbres.

Hyperparamétrage de la forêt aléatoire

Le paramétrage du modèle est une étape importante, car les performances de l'algorithme fermer en fonction du paramètre sélectionné. Nous présentons ci-dessous les paramètres qui affectent le plus le modèle.

— Nombre d'arbres : pour obtenir une prévision stable et robuste, un très grand nombre d'arbres est nécessaire. Mais le temps de calcul augmente avec le nombre d'arbres qui peuvent faire l'algorithme est très lent. Une façon de résoudre ce problème est de choisir un grand nombre d'arbres pour créer B et arrêter le processus de génération d'arbres lorsque l'erreur se produit Out-of-bag ne tombe pas assez.

— La dimension de la sélection de la création de l'arbre K : la dimension de l'échantillon pour bootstrap est un paramètre à considérer. Un équilibre entre l'adéquation des données pour l'arbre à créer et l'indépendance des arbres.

— Nombre de variables à créer Arbre l : le nombre de variables à déterminer une meilleure coupe est choisie pour éviter le sur-échantillonnage ou Sous-échantillonnage. En fait, si l = n, nous serons confrontés au problème du sur-échantillonnage. Mais si L = 1, alors l'algorithme est complètement aléatoire.

— Profondeur maximale des arbres : si la profondeur des arbres est très grande, alors l'échantillon parfaitement répliqué par l'arbre (sur-échantillonnage). Si cette profondeur est définie sur 1, alors pour chaque arbre, la racine est divisée en deux sous-codes et la prédiction pour chaque région correspond à la moyenne des valeurs observées pour chaque nœud.

2.5 Analyse du rendement et Evaluation des performances

Il est souvent plus difficile d'évaluer les performances d'un algorithme de classification que d'évaluer les performances d'un algorithme de régression. Cependant, différentes méthodes permettent de mesurer les performances de l'algorithme. Le premier concerne la matrice de confusion. L'idée derrière cette matrice est de faciliter la visualisation des performances et mesurer le pourcentage de personnes mal classées. Plus ce pourcentage est élevé, plus nous pouvons s'interroger sur l'efficacité et les performances de l'algorithme. Les lignes de cette matrice se trouvent les classes prédites, tandis que chaque colonne représente une classe réelle. Une matrice de confusion. pour un classificateur binaire est illustré à la figure 5. Sur la diagonale principale de la Matrice sont les cas correctement classés (TN et TP), toutes les autres cellules de la matrice contiennent des exemples mal classés. À partir de cette matrice, nous pouvons calculer diverses mesures. Le plus célèbre est l'Accuracy (la précision), la mesure du nombre d'individus correctement classées à partir du nombre total d'individus.C'est donc la somme des éléments de la Diagonale de la matrice divisée par la somme de la matrice. [17][18]

Accuracy : Accuracy $= \frac{TP+TN}{(TP+FP+TN+FN)}$

		Classe réelle	
		-	+
Classe prédite	-	True Negatives *(vrais négatifs)*	False Negatives *(faux négatifs)*
	+	False Positives *(faux positifs)*	True Positives *(vrais positifs)*

TABLE 2.1 – Illustration d'une matrice de confusion

Cette Matrice donne beaucoup d'Informations, mais parfois, nous préférons une plus courte. C'est pourquoi la Precision d'un classificateur est introduite ici, qui est mesurée comme suit :

Précision : $Precision = \frac{TP}{(TP+FP)}$

où TP est le nombre de vrais positifs et FP est le nombre de faux positifs. La précision sera donc égale à 1, ce qui est un score parfait si le classificateur ne prédit que du Vrai Positif. Cette Précision est souvent utilisée en combinaison avec une autre mesure :

Recall : $Recall = \frac{TP}{(FN+TP)}$

où FN indique le nombre de faux négatifs. Recall mesure le rapport des cas positifs correctement détectés par L'algorithme. Dans certains cas, il peut être préférable de détecter les cas indésirables plutôt que de détecter les cas souhaités. Dans une Situation où un algorithme de prévision du cours des actions est créé, il est considéré comme préférable de maximiser la Précision plutôt que Recall, car il semble plus logique d'atténuer les pertes de transactions plutôt que de sauter à chaque occasion[17].

2.6 Conclusion

Dans ce chapitre, nous avons commencé par déterminer ce qu'est l'apprentissage automatique, ensuite, nous avons parler des différents types d'apprentissage et leurs diffrents algorithmes ainsi que leur mecanismes. Nous avons parler aussi des methodes d'evalution des performance de ces dernier.

Chapitre 3

Prédiction des maladies cardiaques par les techniques d'apprentissage automatique

3.1 Introduction

Les maladies cardiaques sont aujourd'hui l'une des causes de mortalité les plus importantes dans le monde. La prédiction des maladies cardiovasculaires est un défi critique dans le domaine de l'analyse des données cliniques. L'apprentissage automatique s'est avéré efficace pour aider à prendre des décisions et des prédictions à partir de la grande quantité de données produites par le secteur de la santé. Dans ce chapitre, nous visons à développer un modèle basé sur le ML pour détecter les maladies cardiaques. Pour cela nous avons étudié quelques travaux de recherche existants, puis nous avons développé des modèles de prédiction de cette maladie basée sur les algorithmes KNN, RF, DT et LR enfin nous avons optimisé un modèle de ML basé sur le KNN.

3.2 Problématique

Le diagnostic de maladies cardiaques est une tâche difficile, il doit être effectué avec précision, perfection et efficacité car une petite erreur peut causer la mort de la personne. Avec l'augmentation rapide de la puissance de calcul, la disponibilité des données de santé et l'apparition de l'apprentissage automatique, la prédiction de maladies cardiaques est devenue possible. L'apprentissage automatique lorsqu'il est appliqué efficacement, peut aider les médecins à poser des diagnostiques presque parfaits, à choisir les meilleurs médicaments

pour leurs patients et à améliorer la santé générale des patients tout en réduisant les coûts. Dans ce travail nous visons à évaluer la capacité de quelques algorithmes d'apprentissage automatique supervisé tel que KNN, RF, DT, LR pour prédire les maladies cardiaques et à optimiser un modèle de ML basé sur le KNN.

3.3 Etat de l'art

Au cours des dernières années, les chercheurs ont largement développé divers modèles pour détecter les maladies cardiaques d'une personne .Dans cette section, nous avons expliqué certaines des solutions existantes.

Aanshi Gupta et al. [29] ont développé un modèle basé sur le ML pour détecter les maladies cardiaques. Dans ce travail, l'algorithme KNN se distingue comme le meilleur algorithme par rapport à d'autres algorithmes tels que Random Forest, Decision Tree, Support Vector Machine et Naïve Bayes. De plus, un prototype est développé pour valider les résultats. Le prototype consistait en un ensemble de capteurs pour surveiller la santé d'une personne. Il est enfin prédit si une personne est sujette à une maladie cardiaque ou non sur la base du modèle formé précédemment. Ainsi, cette solution fournit non seulement un avantage humain important, mais permet également de fournir des données de surveillance proactive de la santé avec une précision de prédiction de 88,52 %.

Archana Singh et al. [31]Dans cet article, les auteurs ont calculé l'exactitude des algorithmes d'apprentissage automatique pour prédire les maladies cardiaques, ces algorithmes sont le KNN, DT, LR (régression linaire) SVM en utilisant le dataset du référentiel UCI pour l'entrainement et les tests.Après avoir appliqué l'approche d'apprentissage automatique pour l'entrainement et les tests, les auteurs ont constaté que l'exactitude du KNN est beaucoup mieux que celle d'autres algorithmes. L'exactitude est calculée à l'aide de la matrice de confusion de chaque algorithme. L'exactitude de KNN atteint 87 %.

Harshit Jindal et al.[30] L'objectif de ce projet est de vérifier si le patient est susceptible de recevoir un diagnostic de maladie cardiaque cardiovasculaire en fonction de ses caractéristiques médicales telles que le sexe, l'âge, les douleurs thoraciques, le taux de sucre à jeun, etc. Un ensemble de données est sélectionné dans le référentiel UCI avec les

antécédents médicaux et les attributs du patient. En utilisant cet ensemble de données, les auteurs ont prédit si le patient peut avoir une maladie cardiaque ou non. Pour prédire cela, ils utilisent 14 attributs médicaux d'un patient et le classent si le patient est susceptible d'avoir une maladie cardiaque. Ces attributs médicaux sont entraînés sous trois algorithmes : Régression logistique, KNN et Random Forest Classifier. Le plus efficace de ces algorithmes est KNN qui a donné un taux d'exactitude de 88,52 %. Et, enfin, ils classent les patients qui risquent de contracter une maladie cardiaque ou non et cette méthode est également totalement rentable.

Md. Nahiduzzaman et al. [32] Dans cette étude, les auteurs ont proposé deux classificateurs. L'un est un réseau de neurones Perceptron multicouche (MLP) et un autre est une machine à vecteurs de support (SVM). Le travail consiste à classer les maladies cardiaques à deux classes et à cinq classes. Ils ont utilisé la base de données en ligne sur les maladies cardiaques de Cleveland qui se compose de 303 instances avec 5 classes et 13 attributs. Pour le problème de classification à deux classes, SVM a un taux d'exactitude de 92,45 % tandis que la précision de MLP est de 90,57 %. Pour le problème de classification à cinq classes, MLP a une exactitude de 68,86% tandis que SVM est de 59,01%.

Références	Années	Algorithmes utilisés	Exactitude (Accuracy)
[29]	2019	KNN, RF, DT, SVM, NB	KNN à 88,52%
[32]	2019	MLP, SVM	- Classification à 2 classes : SVM à 92,45 % - Classification à 5 classes : MLP à 68,86%
[31]	2020	KNN, DT, LR, SVM	KNN à 87 %
[30]	2021	KNN, RF, Régression logistique	KNN à 88,52

TABLE 3.1 – Comparaison des travaux étudiés.

3.4 Approche et solution proposée

La solution proposée (Figure 3.1.) comprend des étapes, tel que la première étape est désignée comme la collecte des données, la deuxième étape est l'analyse exploratoire des donnéesqui sert à comprendre au maximum les données dont nous disposons pour définir une stratégie de modélisation, la troisième étape est le prétraitement des données.Le prétraitement des données traite les valeurs manquantes, le nettoyage des données et la normalisation en fonction des algorithmes utilisés.Après le prétraitement des données, le classificateur est utilisé pour classer les données prétraitées. Le classificateur utilisédansle modèle proposé est KNN, Decision Tree, Random Forest. Enfin, le modèle proposé est entrepris, où nous avons évalué notre modèle en utilisant la métrique de performance l'exactitude.

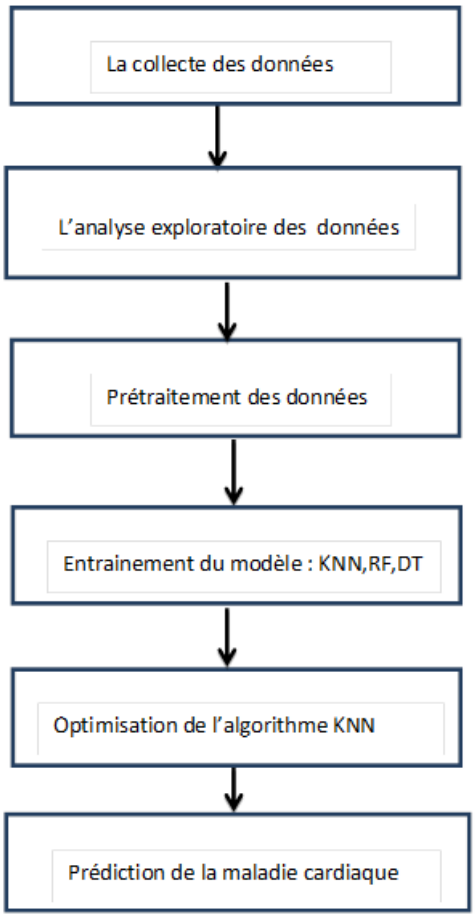


FIGURE 3.1 – Architecture de notre projet

3.4.1 Collecte des données

On a utilisé pour ce travail la base de données de Cleveland qu'est considérée comme un standard pour le système de diagnostic des maladies cardiaques. Elle est tiré du référentiel d'apprentissage automatique de l'UCI [28]. Il y a 303 tuples et 76 attributs dans la base de données. Mais les chercheurs n'ont utilisé que 14 attributs pour le diagnostic des maladies cardiaques. Les étiquettes de classe se composent de deux valeurs telles que 1 pour le patient normal, 0 pour le patient anormal. 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 :masculin, 1 :féminin)	catégorique
3	Cp	Le type de douleur thoracique est classé en quatre types de douleur thoracique selon le patient soit une douleur causée par une angine typique ou causée par une angine atypique ou une douleur non bouchante et peut également apparaître 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	Resting	Résultat électrocardiographique au repos (0 à 1)	catégorique
8	Thali	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 excerice ou tout stress physique ou mental (oui :1, non :0)	catégorique
10	Oldpeak	ST dépression induite par l'exercice par rapport au ('ST'se rapporte aux positions sur le graphique ECG.)	numérique
11	Slope	la pente du segment peak exercise ST (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	Maladie cardiaque (1 ou 0)	catégorique

TABLE 3.2 – Les attributs de dataset

3.4.2 L'analyse exploratoire des données

L'objectif de cette étape est de comprendre au maximum les données dont nous disposons pour définir une stratégie de modélisation.

L'analyse de la forme

Pour faciliter le processus de l'analyse exploratoire des données, nous utilisons la bibliothèque « pandas-profilage » peut générer des rapports interactifs à n'importe quel ensemble de données, avec une seule ligne de code.[26][27]
La sortie est enregistrée en tant que rapport HTML (voir la figure ci-dessous)

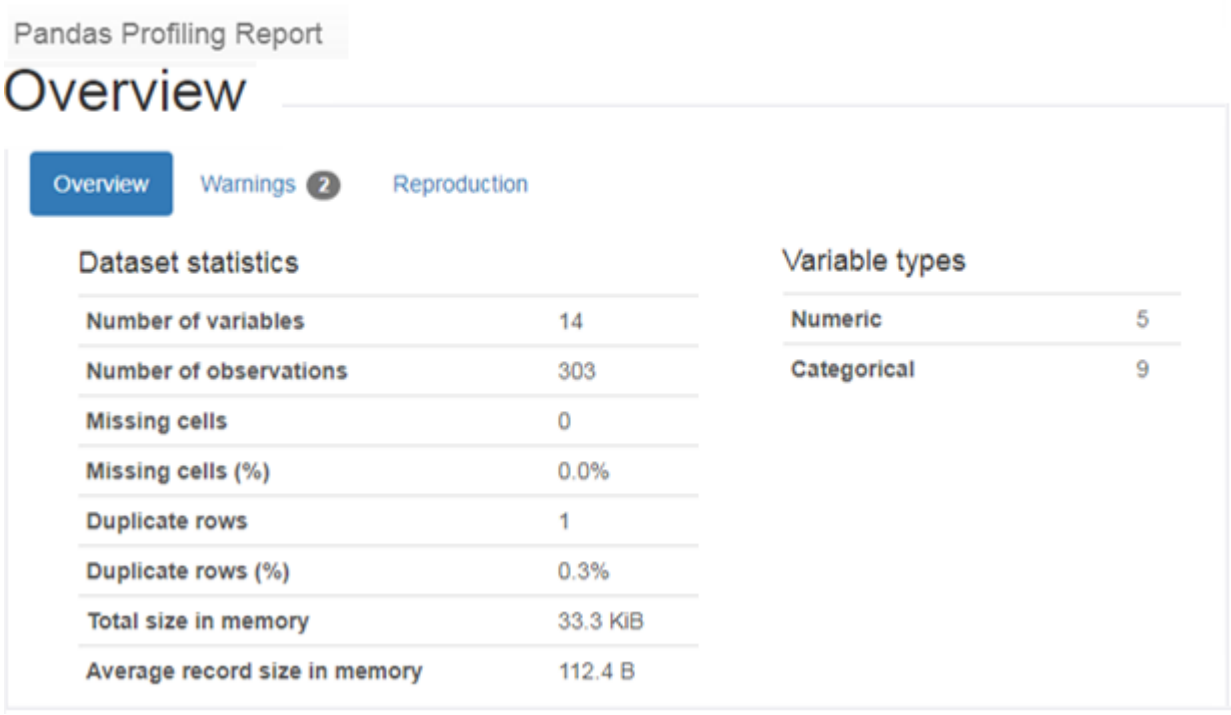


FIGURE 3.2 – Rapport HTML de dataset

Ce rapport nous permet de lire facilement des informations de base sur l'ensemble de données telles que :

— Nombre de variables 14 , 5 numérique et 9 variable Catégorique (variable résultat) .
— Le nombre des observations (303 patient).
— Cellules manquantes 0 cellule.
— Répétitions de ligne nous avons 1 répétition avec un de pourcentage est 3%.

— La taille de l'ensemble de données.

— Taille moyenne de l'enregistrement en mémoire.

L'analyse du fond

Visualisation de la classe cible :

L'équilibrage des données est essentiel pour un résultat précis, grâce au graphique d'équilibrage des données, nous pouvons voir que les deux classes cibles sont égales. La Figure 3.3 représente les classes cibles.

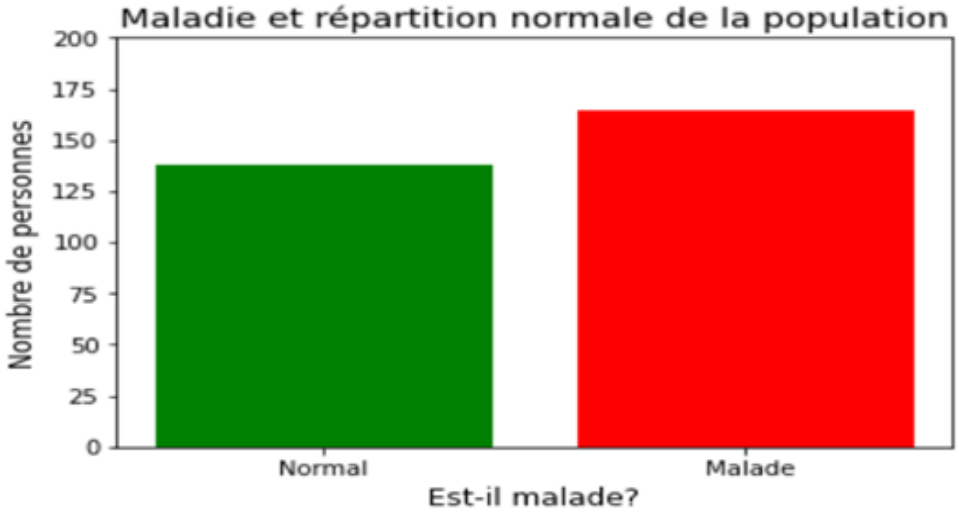


FIGURE 3.3 – Visualisation de la classe cible

Le graphique ci-dessus montre que le nombre de personnes sans maladie cardiaque est de 138 avec une proportion de 45,54% et le nombre de personnes atteintes de maladies cardiaques est de 165 avec une et la proportion de 54,46%. Nous notons que le pourcentage de personnes atteintes de maladies cardiaques et de personnes sans maladie est proche, ce qui évite les problèmes de déséquilibre des données.

Visualisation de la relation variables/classe cible

— **Maladie cardiaque/âge**

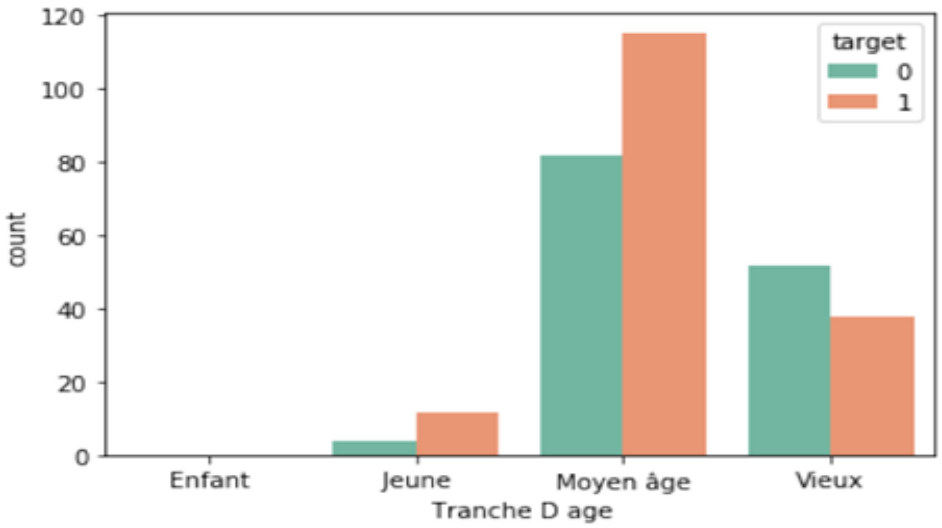


FIGURE 3.4 – Visualisation relation âge /maladie cardiaque

Les résultats montrent que la populationd'âge moyen est la plus touchée par cette maladie.

— **Relation maladie cardiaque/sexe**

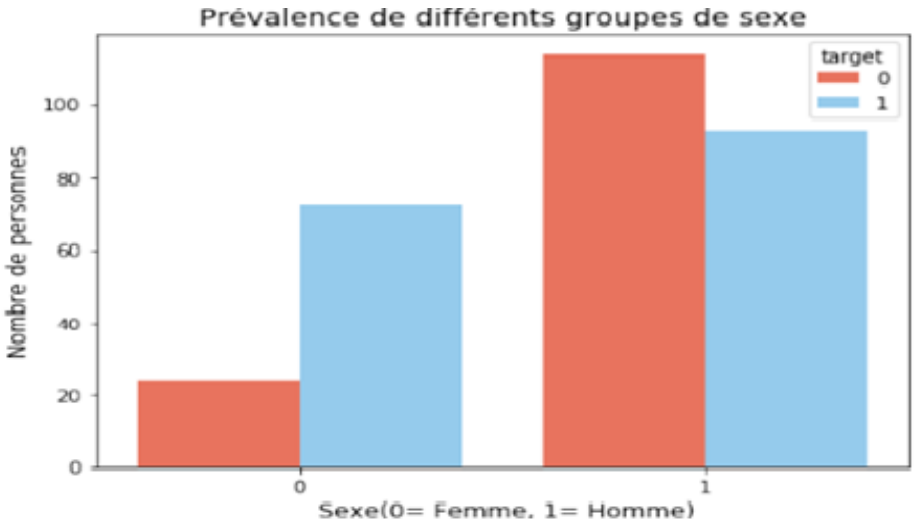


FIGURE 3.5 – Relation maladie cardiaque/sexe

Les résultats montrent que la proportion de femmes souffrant de maladies cardiaques est plus élevée que celle des hommes.

— **Relation maladie cardiaque/ type de douleurs thoracique (Cp)**

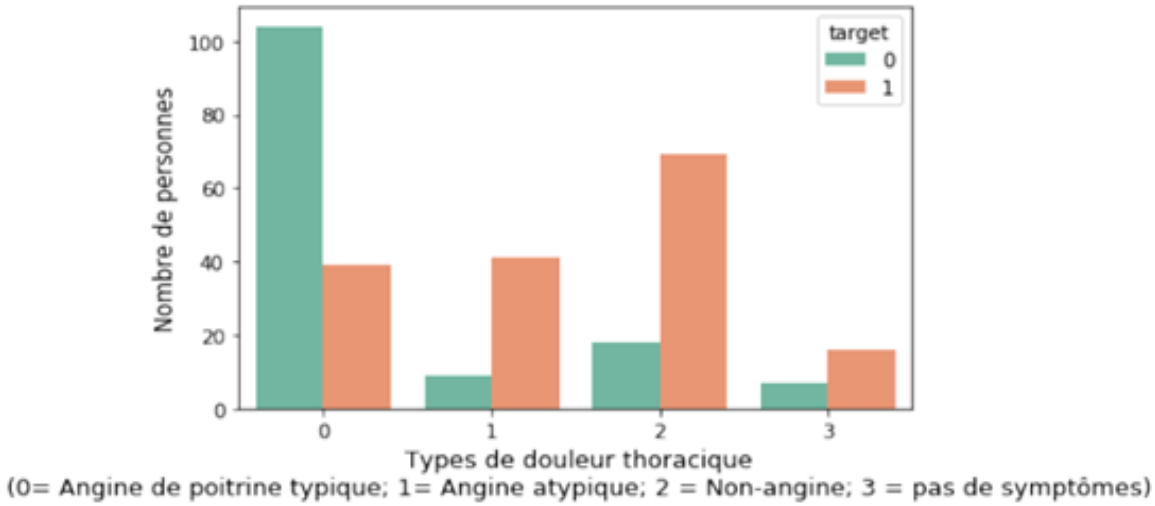


FIGURE 3.6 – Relation maladies cardiaque/Cp

Les résultats montrent que la probabilité de maladie cardiaque chez les patients souffrant d'angine de poitrine typique est relativement faible. Il ne s'agit que d'une angine pure et non d'une maladie cardiaque, tandis que d'autres patients souffrant d'angine de types 2, il ont un taux plus élevé de maladie cardiaque, ce qui indique une certaine relation entre les maladies cardiaques et le type d'angine de poitrine

— **Relation maladie cardiaque/fréquence cardiaque maximale (thalach)**

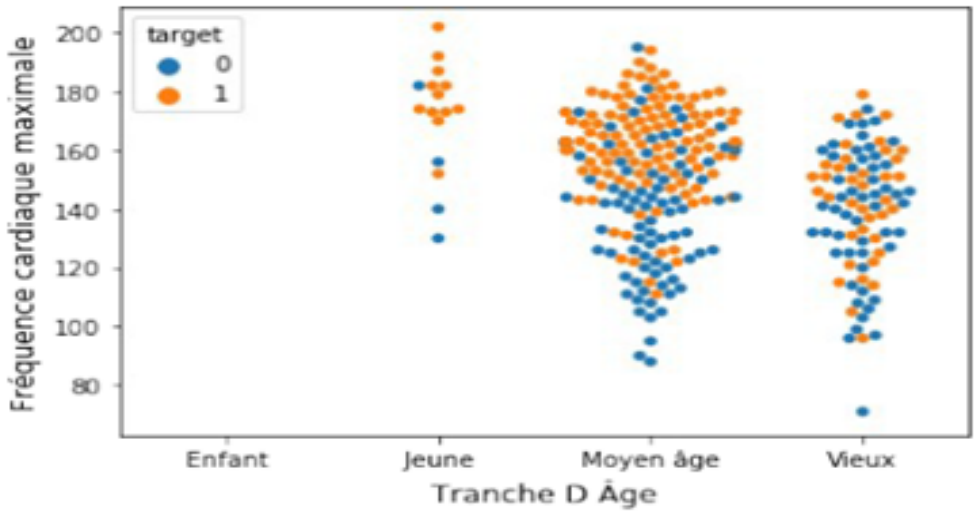


FIGURE 3.7 – Relation maladie cardiaque/thalach

Les résultats montrent que :

1. À mesure que l'âge augmente, la fréquence cardiaque maximale diminue progressivement.
2. Dans le même groupe d'âge, la fréquence cardiaque des personnes atteintes d'une maladie cardiaque est généralement plus élevée que celle des personnes normales.

— **Relation maladies cardiaque/la pression artérielle au repos**

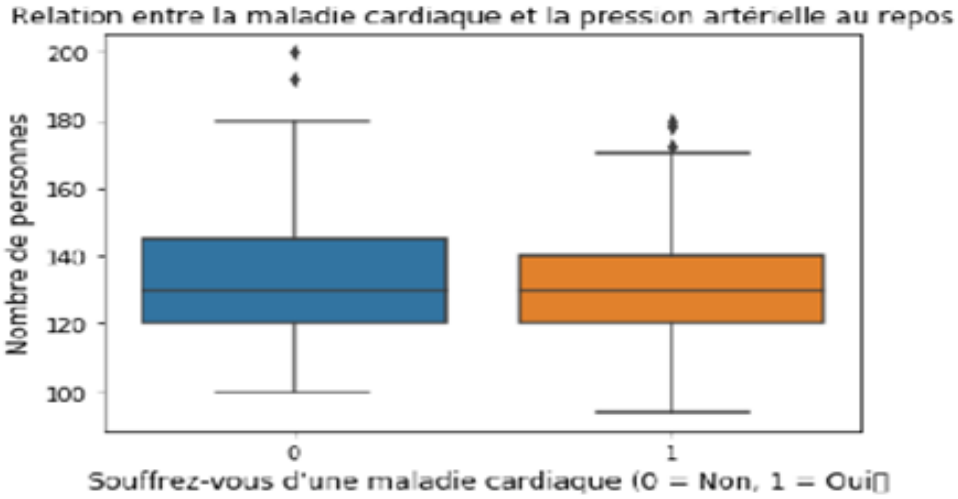


FIGURE 3.8 – Relation maladie cardiaque/la pression artérielle au repos

Les résultats montrent que la pression artérielle au repos des personnes normales est légèrement supérieure à celle des patients atteints de maladies cardiaques.

La sélection d'attributs

Le chef professionnel choisit ses matières premières avec soin afin d'avoir un produit de bonne qualité, il en va de même pour l'apprentissage automatique algorithmique, chaque fois que les données sont excellentes, les algorithmes nous donneront d'excellents résultats. L'un des processus importants pour améliorer les données est de déterminer la corrélation entre les variables, où elle est utilisée pour représenter la mesure statistique de la relation linéaire entre deux variables. Elle peut également être définie comme une mesure de dépendance entre deux variables différentes. Pour calculer le rapport de corrélation entre

deux variables différentes, nous avons utilisé le coefficient de corrélation de Pearson défini par la formule ci-dessous.

$$\frac{\sum (x_i - \overline{x}) * (y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 * \sum (y_i - \overline{y})^2}}$$

Les valeurs que nous obtenons du coefficient de corrélation sont limitées entre -1 et 1

- Si la valeur est -1, nous dirons qu'il s'agit d'une corrélation négative entre deux variables. Cela signifie que lorsqu'une variable augmente, l'autre variable diminue.
- Si la valeur est 0, il n'y a pas de corrélation entre deux variables. Cela signifie que les variables changent de manière aléatoire les unes par rapport aux autres.
- Si la valeur est 1, nous dirons qu'il s'agit d'une corrélation positive entre deux variables. Cela signifie que lorsqu'une variable augmente, l'autre variable augmente également.

Les caractéristiques corrélées ne sont pas utiles pour la variable prédictive car elles ajoutent une complexité du calcul. Nous avons utilisé la matrice de corrélation pour vérifier si les attributs sont corrélés ou non. La matrice de corrélation expérimentée est représentée sur la figure 3.9. Nous observons que la corrélation entre deux variables ne dépasse jamais 0.45, ce qui est acceptable. Une corrélation supérieure à 0,7 peut poser problème et ces caractéristiques doivent être éliminées. Les cases jaune en diagonale dans la matrice de corrélation montrent que les caractéristiques sont complètement corrélées à elles-mêmes, ce qui est vrai car elles ont une corrélation de 1. Pour notre cas, nous n'éliminons aucune variable et le jeu de données prétraité final comprenait 14 variables. Voici une représentation graphique de la matrice de corrélation de l'ensemble de données sur les patients cardiaques avec le code python pour la générer. [24][25]

```
import matplotlib.pyplot as plt
plt.figure(figsize=(12,10))
sns.heatmap(data.corr(),annot=True,cmap="magma",fmt='.2f')
```

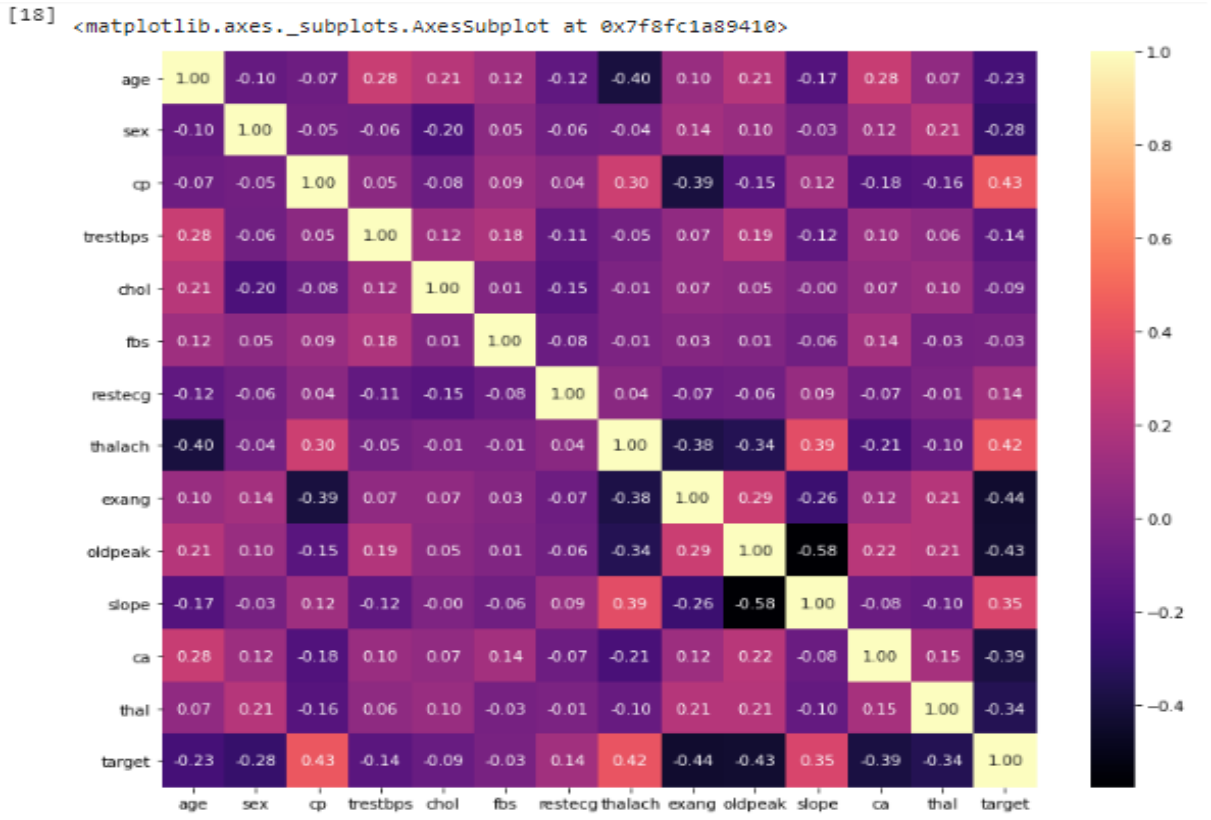


FIGURE 3.9 – Matrice de corrélation

3.4.3 Prétraitement de données

L'objectif de l'étape de prétraitement de données est de transformer l'ensemble des données pour le mettre dans un format propice au ML, car les données brutes sont souvent déformées et peu fiables, et elles peuvent y avoir des valeurs manquantes.

Le remplacement des valeurs manquantes :

Il existe plusieurs façons pour résoudre le problème de certaines valeurs manquantes, l'une des façons les plus simples pour résoudre ce problème consiste simplement à supprimer les lignes qui possèdent des valeurs manquantes. Pour notre base de données «heart .csv », nous avons constaté qu'elle ne possède pas de valeurs manquantes comme mentionné dans la figure ci-dessous.[26][27]

```
#Connaître les valeurs vides dans chaque colonne
data.isnull().sum()
```

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

FIGURE 3.10 – Les valeurs manquantes

On peut rechercher les valeurs manquantes en utilisant la fonction du Heatmap En remarque qu'il n y a de valeur manquante.

FIGURE 3.11 – Recherche de valeurs manquantes à l'aide de Heatmap

Suppression des doublons :

Les valeurs en double affectent négativement les résultats de la prédiction, et nous pouvons observer que l'algorithme KNN, par exemple, si nous voulons prédire une valeur proche d'une valeur en double, cela signifie qu'on va prendre la même classe que la valeur

en double. NumPy et Pandas offrent des moyens simples pour supprimer les lignes en double. Notre ensemble de données contient 302 valeurs non dupliquées et une valeur dupliquée.

```
#répétitives sont "False" et répétitions est "True"
print(data.duplicated().value_counts())
```

```
False    302
True       1
dtype: int64
```

Suppression des lignes doublée

```
#Supprimer les valeurs répétées
data=data.drop_duplicates(keep='first')
print(data.duplicated().value_counts())
```

```
False    302
dtype: int64
```

Mise à l'échelle des Données (Data Scaling)

L'une des étapes les plus importantes consiste à traiter les préréglages de données appliqués à des variables indépendantes ou à des entités de données. Cela aide essentiellement à normaliser les données dans une certaine plage. Parfois, cela aide également à accélérer les calculs dans un algorithme, en particulier comme l'algorithme du voisin le plus proche, dont les étapes sont le calcul de distance.[22] L'une des techniques les plus importantes pour effectuer une mise à l'échelle qui est la plus utilisée c'est « Standardisation ». C'est une technique très efficace qui redimensionne une valeur de caractéristique de sorte qu'elle ait une distribution avec une valeur moyenne 0 et une variance égale à 1.[23] Le point de normalisation est de modifier les observations afin qu'elles puissent être décrites comme une distribution normale.

— **Distribution normale** : Également connue sous le nom de" courbe en cloche", il s'agit d'une distribution statistique spécifique où des observations à peu près égales se situent au-dessus et en dessous de la moyenne, la moyenne et la médiane sont les mêmes, et il y a plus d'observations plus proches de la moyenne. La distribution normale est également connue sous le nom de distribution gaussienne

```
from sklearn.preprocessing import StandardScaler
sc_X = StandardScaler(copy=True)
X = sc_X.fit_transform(X)
```

Fonction mathematique populaire aussi utilise par sklearn

$$X_{new} = \frac{X_i - X_{mean}}{StandardDeviation}$$

3.4.4 Entrainement de modèle

Cette étape consistait à diviser l'ensemble de données prétraité en un ensemble de données d'entraînement et de test. Nous avons choisi 77% pour les données d'entrainement et 23% pour les données de test puis nous entamerons l'entraînement du modèle à l'aide de l'ensemble de données d'entraînement en utilisant les algorithmes KNN, DT, RF. Ensuite, les prédictions sont faites à l'aide de l'ensemble de données de test sur différents modèles entraînés (avec différents algorithmes). À partir des tests, nous avons le meilleur modèle sélectionné qui est le KNN avec une valeur de K= 7 en utilisant la distance euclidienne en fonction de la mesure de performance l'exactitude qui vaut 91.42%. Ce modèle est sélectionné comme modèle d'entraînement final pour d'autres prédictions.

3.4.5 Optimisation de l'algorithme KNN

Dans le chapitre 2 nous avons expliqué le principe de fonctionnement de l'algorithme KNN et nous avons mentionné ces avantages et ces inconvénients. Pour optimiser et augmenter l'efficacité de l'algorithme KNN, nous devons d'abord trouver les cas erronés qui minimise l'efficacité de l'algorithme.

Les cas erronés de l'algorithme KNN

Cas 1 :

FIGURE 3.12 – Le premier cas erroné du KNN

Dans ce cas, si le nombre de voisins est supérieur ou égal à 7, l'algorithme KNN prend les cas positifs, et ignore la population négative malgré qu'il soit proche de cette dernière.

Cas 2 :

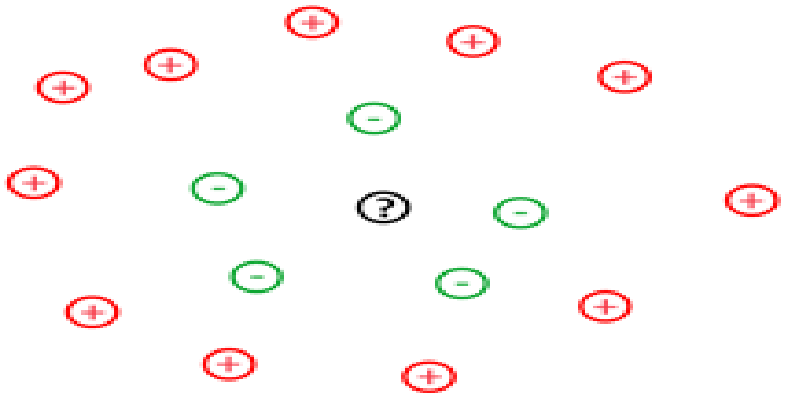


FIGURE 3.13 – Le deuxième cas erroné prévu par le KNN

Dans le même cas que le dernier, si le nombre de voisins est inférieur à 11, le point non classifié sera considéré soit un cas négatif ou un positif. Comme nous le constatons, le cas testé est au milieu d'un groupe des cas négatifs, ici le KNN ne prend pas en compte les petits groupes, et cherche ailleurs, comme nous le voyons les cas négatifs forment un groupe très proche par rapport aux autres.

Cas 3 :

FIGURE 3.14 – Le troisième cas erroné prévu par le KNN

Dans ce cas, si nous testons un point non classifié qui a les mêmes coordonnées qu'un point négatif et que K est égal à 3, alors l'algorithme KNN régulier, trouve que parmi ses voisins se trouve un point négatif qui a les mêmes coordonnées avec lui et a 2 points voisins positifs, alors il considère que l'état de dernier est positif.

Cas 4 :

FIGURE 3.15 – Le quatrième cas erroné prévu par le KNN

Parmi les problèmes les plus commun est que l'algorithme KNN fait face à un nombre égale de cas de deux catégories différentes , la-il choisie aux hasard

3.4.5.2. La solution proposée

La solution proposée consiste à optimiser l'algorithme du voisin le plus proche, au lieu de prendre la décision à partir de la population dominante, nous prendrons la décision partir d'un nouveau calcul de probabilité et la distance.Dans l'algorithme du voisin le plus proche, l'algorithme calcule la distance entre l'état testé et toutes les données d'apprentissage, puis identifie les K plus proches voisins, puis parmi les K voisins, il détermine la classe de l'état testé à partir de la classe d'états dominantes Dans l'algorithme KNN amélioré, nous calculons les distances avec toutes les données d'apprentissage et sélectionnons les

voisins les plus proches comme dans l'algorithme normal, mais au lieu de décider à partir de la classe dominante, nous effectuons des opérations sur les distances des voisins les plus proches, nous multiplions donc chacune des distances que nous avons par la probabilité de la classe a qui appartient à Pour cela, où nous obtenons une nouvelle distance, cette nouvelle distance que nous soustrayons de la distance d'origine comme indiqué dans la formule qui suivant :

$$D_{Nouveau} = D_{Précédent} - (D_{Précédent} * (p_{La\ probabilité\ de\ chaque\ voisin}))$$

$D_{Nouveau}$: La distance qui permet de classer le cas non classé

-$D_{Précédent}$: C'est la distance que nous avons obtenue dans la première étape pour déterminer les points adjacents.

-$\mathbf{p}_{La\ probabilité\ de\ chaque\ voisin}$: C'est la probabilité de la même classe que le point adjacent pour l'état non classé.

Après avoir calculé les nouvelles distances, nous les organisons et attribuons la catégorie du cas testé à partir de la même catégorie de la distance la plus basse, Nous donnerons des exemples pour illustrer l'idée plus.

3.4.5.2.1. Résolution des cas erronés du KNN avec le KNN optimisé

— **Premier exemple expilactif :** Supposons que K = 3 et que les voisins soient comme l'indique la figure, nous avons dessiné des cercles illustratifs dont le centre est le point voisin et son rayon qui est eagale :

$$(D * p)$$

— **D** : La distance entre le point non classé et le point classé, C'est la distance du point adjacent qui est le voisin le plus proche dans la première étape de l'algorithme KNN

— **p** : La probabilité d'un point adjacent est le nombre de points de la même classe divisé par le nombre de voisins K.

Le travail que nous avons fait est de calculer la probabilité de chaque classe par rapport au point non classé, et nous multiplions la distance entre le point non classé et le point adjacent par sa probabilité de classe, nous obtenons une distance qui est inférieure à la première distance de parcours, nous allons annuler cette distance qui ne se pas prise en considération car l'important c'est la distance qui reste.

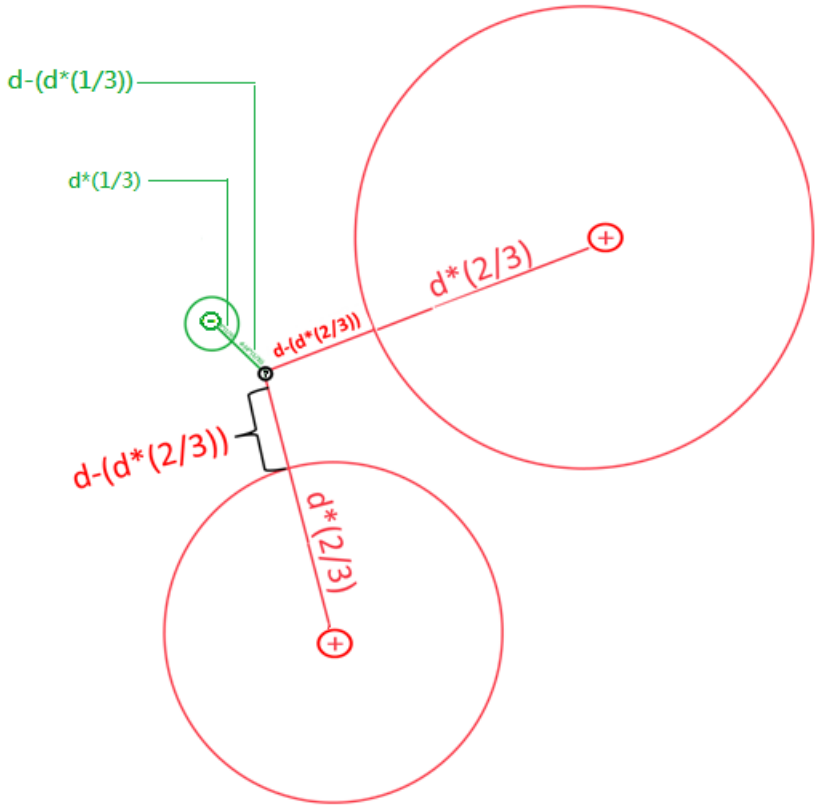


FIGURE 3.16 – exemple1 d'un cas erroné prévu par le KNN

— **Deuxième exemple explicatif** Dans le cas suivant, pour le cas non classé, la probabilité d'un cas positif est de 2/3 et la probabilité d'un cas négatif est de 1/2. Notez que la distance entre le point non classé est égale à celle des points 1 et 2, mais la probabilité du point 1 est supérieure après avoir calculé les nouvelles distances, alors nous admettons que le point 1 est le point le plus proche, et nous dirons que la catégorie du point non classé est un cas positif.

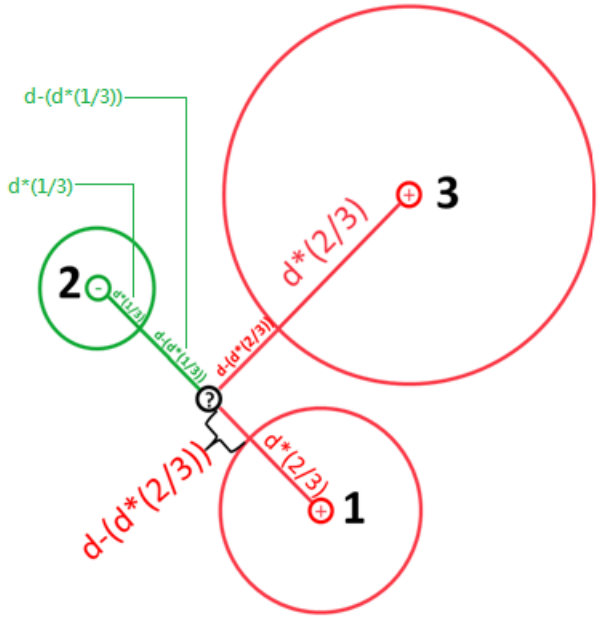


FIGURE 3.17 – exemple2 d'un cas erroné prévu par le KNN

— **Troisième exemple explicatif** Dans ce cas, K = 2 , ce n'est pas recommandable dans l'algorithme normale , pour éviter le problème d'égalité entre deux classes, dans ce cas la probabilité de chaque classe pour le cas non classé est de 1/2, alors la distance entre le point non classés et les deux points adjacents diminue de moitié, et nous dirons que la catégorie du cas non classé est un cas négatif car il est plus proche du négatif.

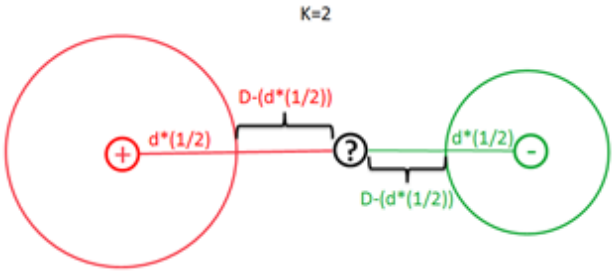


FIGURE 3.18 – exemple3 d'un cas erroné prévu par le KNN

Pour donner plus de flexibilité à l'algorithme, nous ajoutons une valeur appeler Epsilon ε , Le signe sera positif si la probabilité d'une catégorie est la plus grande et prend un signe négatif si c'était le contraire et qu'ils ne lui ont donné aucune valeur, il prend la valeur 0 automatique.

La valeur d'epsilon est ajoutée à chaque cellule, qui est l'augmentation et la diminution du diamètre des cellules, ou en d'autres termes l'augmentation ou la diminution de la longueur de la distance que l'on enlève de la distance d'origine par rapport à la probabilité.

Donc, l'algorithme du voisin le plus proche nous donne les distances entre le cas testé et le voisin le plus proche, nous appliquons la loi aux distances pour obtenir de nouvelles distances, puis nous choisissons la distance la plus courte et la catégorie à laquelle elle appartient, La loi est : Si la probabilité de catégorie est la plus grande, la loi est la suivante :

$$\mathbf{D}_{Nouveau} = \mathbf{D}_{Précédent} - (\mathbf{D}_{Précédent} * (\mathbf{p}_{La\ probabilité\ de\ chaque\ voisin}() + \varepsilon))$$

Si non :

$$\mathbf{D}_{Nouveau} = \mathbf{D}_{Précédent} - (\mathbf{D}_{Précédent} * (\mathbf{p}_{La\ probabilité\ de\ chaque\ voisin}() - \varepsilon))$$

- $\mathbf{D}_{Nouveau}$: La nouvelle distance qui permet de classer le cas non classé.

-$\mathbf{D}_{Précédent}$: C'est la distance que nous avons obtenue dans la première étape pour déterminer les points adjacents.

-$\mathbf{p}_{La\ probabilité\ de\ chaque\ voisin}$: C'est la probabilité de la même classe que le point adjacent pour l'état non classé .

Les étapes de l'algorithme KNN optimisé

- Étape 1 : Sélectionnez le nombre K de voisins.
- Étape 2 : Calculez la distance.
- Étape 3 : Prendre les K voisins les plus proches selon la distance calculée.
- Étape 4 : Parmi ces voisins K, calculez la probabilité de tous les points appartenant à chaque catégorie.

$$p(chaquechaque\ piont) = \frac{(Le\ nombre\ de\ voisins\ de\ la\ même\ classe)}{K}$$

- Étape 5 : Calcul de nouvelles distances liées à la probabilité

  -Si la probabilité de catégorie est la plus grande, la loi est la suivante :

$$\mathbf{D}_{Nouveau} = \mathbf{D}_{Précédent} - (\mathbf{D}_{Précédent} * (\mathbf{p}_{La\ probabilité\ de\ chaque\ voisin}() + \varepsilon))$$

Si non :

$$\mathbf{D}_{Nouveau} = \mathbf{D}_{Précédent} - (\mathbf{D}_{Précédent} * (\mathbf{p}_{probabilité\ de\ chaque\ voisin}() - \varepsilon))$$

— Étape 6 : Triez les distances de la plus petite à la plus grande, définissez la plus petite distance et prenez la classe de cas testée du même cas que la classe de plus petite distance.

3.5 Conclusion

Dans ce chapitre nous avons parlé des différents travaux de recherche pour prédire les maladies cardiaques et leurs résultats obtenus on les comparent entre eux, nous avons vue aussi la méthode d'approche pour développer une solution et les concept de l'analyse exploratoire, les techniques de prétraitements des données, afin d'entrainer des modèles d'apprentissage, ensuite nous avons cherché les cas erronés prévus par l'algorithme KNN afin de l'optimiser. Dans le prochain chapitre nous parlerons des outils de développement utilisé et les résultats obtenus des test de quelques algorithmes compare a notre solution.

Chapitre 4

Résultats et évaluation

4.1 Introduction

Après avoir détaillé l'approche et la solution proposée dans le chapitre précédent, nous montrerons dans ce chapitre les outils de développement et le langage de programmation qu'ils nous ont permis d'implémenter notre solution. En outre, d'une autre part, nous allons discuter en détail les résultats obtenues.

4.2 Environnement de développement

Il existe plusieurs logiciels d'apprentissage automatique disponibles sur le marché. Mais l'une des choses qui vous oblige à utiliser un environnement spécifique est la machine que vous utilisez, et nous allons parler de l'environnement que nous avons utilisé, ainsi que des langages de programmation et des bibliothèques.

4.2.1 La plateforme de développement Anaconda

Anaconda est une distribution libre et open source des langages de programmation Python et R appliqué au développement d'applications dédiées à la science des données et à l'apprentissage automatique (traitement de données à grande échelle, analyse prédictive, calcul scientifique), qui vise à simplifier la gestion des paquets et de déploiement. Les versions de paquetages sont gérées par le système de gestion de paquets anaconda. La distribution Anaconda est utilisée par plus de 6 millions d'utilisateurs et comprend plus de 250 paquets populaires en science des données adaptés pour Windows, Linux et MacOS

[16]

4.2.2 La plateforme de développement Google Colab

Google Colab est un des services proposés par la société Google. Il offre aux étudiants ou scientifiques la possibilité d'exécuter du code écrit en langage Python directement depuis son navigateur. Cela évite pour une personne de devoir installer les logiciels nécessaires pour le faire sur son ordinateur. Il est notamment possible d'interagir avec d'autres services Google, comme Google Drive [17].

4.2.3 Le langage de programmation python

Python est un langage de programmation est le plus populaire et largement utilisé, en particulier dans le domaine de la science des données et d'apprentissage automatique, Python se démarque en offrant en particulier un grand nombre de très haute qualité bibliothèques, couvrant tous les types d'apprentissage qui combinent la facilité d'utilisation et d'apprentissage avec le pouvoir des bibliothèques qu'il possède. Parmi ces bibliothèques, nous avons utilisé [18].

4.2.4 Les bibliothèques utilisées

Scikit-learn

Scikit-learn est une bibliothèque libre Python destinée à l'apprentissage automatique. Elle est développée par de nombreux contributeurs notamment dans le monde académique par des instituts français d'enseignement supérieur et de recherche comme Inria [20].

Pandas

Pandas est une bibliothèque écrite pour le langage de programmation Python permettant la manipulation et l'analyse des données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques et de séries temporelles [21].

Seaborn

Seaborn est une bibliothèque de visualisation de données Python basée sur matplotlib. Il fournit une interface de haut niveau pour dessiner des graphiques statistiques attrayants et informatifs [22].

Matplotlib

Matplotlib est une bibliothèque du langage de programmation Python destinée à tracer et visualiser des données sous formes de graphiques [23].

4.3 Résultats et analyse

4.3.1 Evaluation des performances de KNN, RF, DT, LR

Après avoir entraîné les modèles avec les algorithmes KNN, RF, LR et DT, nous prédisons les classes en utilisant l'ensemble de données de test afin de sélectionner le meilleur modèle pour l'utiliser dans la phase de prédiction. Le meilleur modèle est choisi en fonction de ses performances. En général, les performances d'un algorithme de Machine Learning sont évaluées en fonction de paramètres tels que l'exactitude, la précision, le rappel, la spécificité, etc. Dans notre cas nous avons choisi l'exactitude comme critère de performance pour évaluer le modèle d'entraînement.

Dans ce qui suit la matrice de confusion des différents modèles.

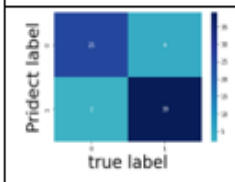
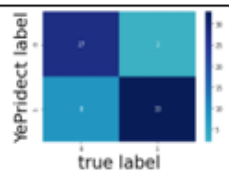
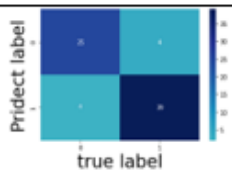
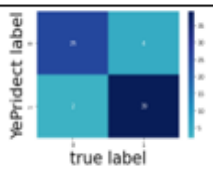
KNN			Decision Tree			Logistic Regression			Random Forest		
	False	True		False	True		False	True		False	True
False	25	4	False	27	2	False	25	4	False	25	4
True	2	39	True	8	33	True	4	37	True	2	39
											

TABLE 4.1 – la matrice de confusion des différents modèles.

Les résultats montrent que parmi les quatre modèles, le KNN et Random Forest ont le taux de vrais positifs le plus élevé, suivie par régression logistique, et l'arbre de décision a le plus bas.

A partir de la matrice de confusion, nous calculons l'exactitude (Accuracy) de chaque algorithme.

Algorithme	**Accuracy**
KNN	91,42%
Decision Tree	85,71%
Logistic Regression	88 ,15%
Random Forest	91,42%

TABLE 4.2 – les métriques de performance de chaque algorithme.

D'après les résultats présentés dans la table ci-dessus, nous notons que l'algorithme KNN et RF avait le même taux de prédiction et était le meilleur parmi tous les algorithmes, atteignant 91,42 %, à partir de ces résultats, nous avons proposé une optimisation de l'algorithme du K plus proche voisin, mais avant cela nous avons essayé d'améliorer le taux de prédiction en préparant bien les données, nous avons donc obtenu les résultats suivants.

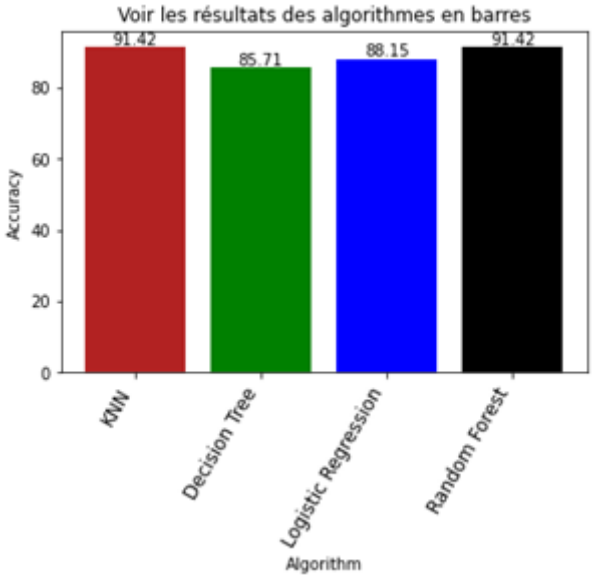


FIGURE 4.1 – La comparaison de la précision des différents algorithmes.

4.3.2 Evaluation des performances de KNN régulier et KNN optimisé.

Après avoir optimisé l'algorithme du plus proche voisin, la première chose que nous voulions savoir est si l'algorithme que nous avons optimisé donne de meilleurs résultats que l'algorithme KNN régulier, mais nous les avons testés avec des valeurs de K= 1 à K=20 et enregistré le taux de prédiction à chaque incrémentation de K et nous avons tracé les résultats dans un graphique pour une analyse plus facile et meilleure, nous avons essayé d'analyser les deux figures comme le montre le tableau suivant :

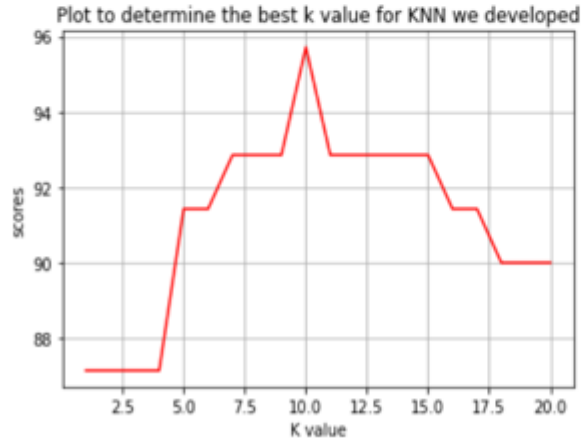
KNN régulier (sklearn)		Dans le graphique suivant, nous remarquons que les résultats de prédiction s'est amélioré dans les trois portées K=5, K = 7 et K = 15 où le rapport de prédiction a atteint 91,42%
KNN optimisé		Dans le graphique suivant de l'algorithme du voisin le plus proche que nous avons optimisé, on note qu'il atteint la valeur la plus élevée à K=10, où le ratio de prédiction a atteint 95,714%

TABLE 4.3 – Tableau d'analyse pour déterminer la meilleure valeur K pour KNN et KNN optimisé

Pour essayer de simplifier l'analyse, nous avons dessiné les deux graphiques dans un seul plan, où l'algorithme KNN que nous avons optimisé est en rouge et l'algorithme KNN normal est en bleu, et nous noterons que l'algorithme que nous avons développé fait un meilleur score que le KNN régulier algorithme pour toutes les valeurs de K, voir le tableau suivant :

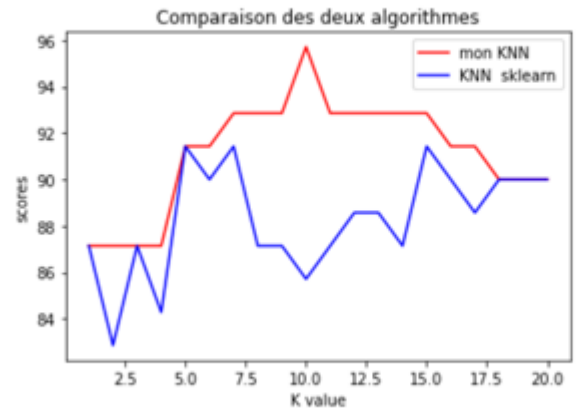
Comparaison des deux algorithmes		Selon ce graphe, nous notons que notre algorithme que nous avons optimisé est meilleur que l'algorithme régulier du plus proche voisin. (KNN sklearn)

TABLE 4.4 – Tableau d'analyse pour déterminer la meilleure valeur K pour KNN régulier et KNN optimisé.

Nous allons maintenant comparer les résultats de notre algorithme avec les résultats des algorithmes précédents et nous allons d'abord voir la matrice de confusion dans le tableau suivant :

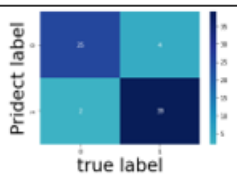
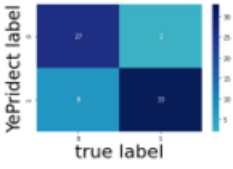
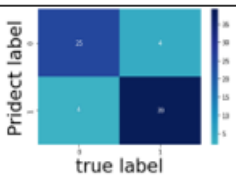
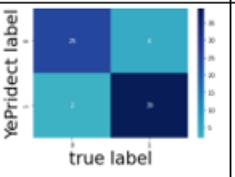
KNN			Decision Tree			Logistic Regression			Random Forest			KNN optimisé		
	False	**True**		**False**	**True**		**False**	**True**		**False**	**True**		**False**	**True**
False	**25**	**4**	**False**	**27**	**2**	**False**	**25**	**4**	**False**	**25**	**4**	**False**	**28**	**1**
True	**2**	**39**	**True**	**8**	**33**	**True**	**4**	**37**	**True**	**2**	**39**	**True**	**2**	**39**
														

TABLE 4.5 – matrice de confusion des algorithme KNN, DT, LR, RF et KNN-optimisé.

Le tableau suivant nous permet de voir plus clairement les résultats de la précision en pourcentage des différents modèles comparer au notre dans le tableau et ci-dessus :

Algorithme	**Accuracy**
KNN	**91,42%**
Decision Tree	**85,71%**
Logistic Regression	**88.15%**
Random Forest	**91,42%**
KNN optimisé	**95,71%**

TABLE 4.6 – comparison de l'exactitude.

Les résultats montrent que parmi les cinque modèles la précision de notre KNN optimise est nettement supérieur aux autre modèles, alors que Random Forest et le KNN régulier ont le même taux de précision , suivie par régression logistique, et l'arbre de décision le taux le plus bas.

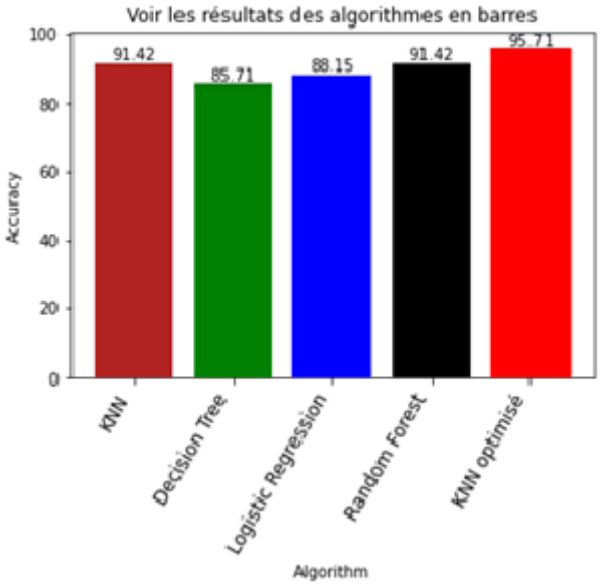


FIGURE 4.2 – Comparaison de l'exactitude pour différents modèles.

Le graphe des résultats obtenue nous montre bien que l'algorithme que nous avons optimise donne des résultat meilleur que les autres algorithmes.

4.4 Conclusion

Dans ce chapitre, nous avons présenté les différents outils et langage de développement qui permettent l'utilisation d'algorithmes, ainsi que la solution que nous avons proposée pour l'optimisation KNN, où nous avons comparé les résultats de prédiction de certains des algorithmes étudiés avec celle de l'algorithme amélioré.

Conclusion générale

Les maladies cardiaques sont considérées par des études, c'est la première cause de décès dans le monde et en Algérie avec un taux de 34%. Par conséquent, un diagnostic précis à un stade précoce suivi d'un traitement ultérieur approprié peut sauver un grand nombre de vies. Ce qui a conduit à l'intervention de la science moderne pour diagnostiquer et prédire les maladies, y compris les maladies cardiaques, à de l'apprentissage automatique. À travers cette thèse, nous espérons avoir atteint nos objectifs, on mesure de prédiction des maladies et d'optimisation de l'algorithme du KNN.

Tout cela nous a conduit à prendre des algorithmes d'apprentissage automatique les plus utilisés, à savoir KNN, RF, LR et DT, et faire avec des prédictions en les comparant entre eux, nous avons parlé de l'analyse des donnes et les différentes manières de manipulations et de nettoyage afin d'avoir des donnes propres et utiles pour nourrir les différents modèle d'apprentissages de bonne performance de ces derniers.

Après plusieurs tests, les résultats, nous montrons que l'algorithme KNN avait le meilleur taux de prédiction parmi le reste des algorithmes, et à partir de ces résultats, nous avons suggéré une amélioration pour l'algorithme KNN. Après avoir exploré les défirent cas où il est susceptible de ne pas bien performer, nous avons proposé une méthode de calcul des distances entre les voisins les plus proches, qui a résolu certain des cas erronés, et cela, après avoir testé avec plusieurs techniques différentes, les résultats ont été satisfaisants, car nous avons enregistré le taux de prédiction le plus élevé en le faisant correspondre avec l'algorithme KNN normal et le reste des algorithmes utilisés. Nous espérons aussi que notre travail sera un plus dans le domaine de l'apprentissage automatique et d'un bénéfice pour ceux qui font leur recherche dans l'amélioration des diagnostiques dans le domaine médical.

Ce mémoire nous a permis de d'explorer des domaines nouveaux, qui nous ont permis d'acquérir des connaissances dans le domaine de l'analyse des donnes et de l'apprentissage automatique, qui ont ouvert nous yeux sur nous connaissance et lacune qui se sont amélioré par la suite.

Perspective

Comme perspectives, nous proposons d'utiliser les techniques d'apprentissage en profondeur (deep learning) pour une meilleure analyse et prédiction précoce des maladies cardiaques afin que le taux de décès puisse être minimisé par la sensibilisation aux maladies, ainsi étendre ce projet pour la surveillance et la prédiction d'autres maladies comme le diabète.

Bibliographie

[1] « Le cœur : son fonctionnement, son rôle et ses maladies ». [https://www.doctissimo.fr/html\linebreak/sante/mag_2001/mag0413/dossier/sa_3819_coeur_pompe.htm] , (Consulté 2021).

[2] ANATOMIE CARDIO-VASCULAIRE (2004), [https://www.infirmiers.com/pdf/Anatomie-cardio-vasculaire.pdf]. (Consulté en 2021).

[3] « Comment fonctionne le cœur ». [https://www.coeuretavc.ca/maladies-du-coeur/qu-est-ce-que-les-maladies-du-coeur/comment-fonctionne-le-cur]. (Consulté en 2021).

[4] « Thierry GAULT, Le Cœur » (Janvier 2013). [http://f2.quomodo.com/5C852034/uploads/8489/physiologie\%20le\%20coeur\%20-\%20gault\%2013.pdf]. (Consulté en mai 2021).

[5] « Les maladies cardiovasculaires, première cause de mortalité en Algérie ».[https://www.aps.dz/sante-science-technologie/119636-les-maladies-cardiovasculaires-premiere-cause-de-mortalite-en-algerie], (Consulté en 2021).

[6] Ahmed Elsherif, Suez Canal University ; Stanley Oiseth, Chief Medical Editor, Lecturio. [https://www.lecturio.com/magazine/lecturio-medical-knowledge-essentials-physical-examination-of-the-cardiovascular-system/#:~:text=Cardiovascular\%20examination\%20consists\%20of\%20assessment,and\%20auscultation\%20of\%20the\%20heart]. (Consulté en 2021).

[7] Walker HK, Hall WD, Hurst JW Clinical Methods : The History, Physical, and Laboratory Examinations. 3rd edition. Chapter 7An Overview of the Cardiovascular System (1990) . [https://www.ncbi.nlm.nih.gov/books/NBK393/] (Consulté en 2020) .

[8] [https://www.webteb.com/articles/\%D9\%83\%D9\%8A\%D9\%81\%D9\%8A\%D8\%A9-\%D8\%A7\%D9\%84\%D9\%88\%D9\%82\%D8\%A7\%D9\%8A\%D8\%A9-\%D9\%85\%D9\%86-\%D8\%A7\%D9\%85\%D8\%B1\%D8\%A7\%D8\%B6-\%D8\%A7\%D9\%84\%D9\%82\%D9\%84\%D8\%A8_346] , (Consulté en 2020).

[9] Y.Bastanlar M.ozuysal ." Introduction to machine learning"." In miRNomics : MicroRNA Biology and Computational Analysis", p105–128. Springer, 2014.

[10] R. S. Sutton and A. G. Barto, « Reinforcement Learning : An Introduction », in IEEE Transactions on Neural Networks, vol. 9, no. 5, pp. 1054-1054, Sept. 1998,

[11] TM Mitchell . « Machine learning. 1997 » - Burr Ridge, IL : McGraw Hill, 1997

[12] Jason Wong.(le 8 Décembre 2020) « Logistic Regression Explained », [https://towardsdatascience.com/logistic-regression-explained-afc267815943]. (Consulté en 2021).

[13] Jason Brownlee (le 1 Avril 2016). « Logistic Regression for Machine Learning ». [https://machinelearningmastery.com/logistic-regression-for-machine-learning/] , (Consulté en Décembre 2020).

[14] J. Brownlee, (le 11 Décembre 2017). « Difference Between Classification and Regression in Machine Learning ». [https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/], (Consulté en Décembre 2020)

[15] Marin Ferecatu (2010), « Apprentissage, réseaux de neurones et modèles graphiques (RCP209) Arbres de décision », [http://cedric.cnam.fr/vertigo/cours/ml2/coursArbresDecision.html]. (Consulté en 2021)

[16] Breiman et al. «Random forests, Machine Learning», (2001).

[17] Géron, A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems. .O'Reilly Media 2019.

[18] Openclassrooms. « Évaluez et améliorez les performances d'un modèle de machine learning », (2020) [https://openclassrooms.com/fr/courses/4297211-evaluez-les-performances-dun-modele-de-machine-learning/4308256-evaluez-un-algorithme-de-classification-qui-retourne-des-valeurs-binaires].

[19] Escofier, Brigitte et Pagès, Jérôme. 2016 - 5ème édition. Analyses factorielles simples et multiples. Paris : Dunod, 2016.

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

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

[22] " Python — How and where to apply Feature Scaling ? . (2020). [https://www.geeksforgeeks.org/python-how-and-where-to-apply-feature-scaling/], (Consulté en 2021).

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

[24] Correlation Concepts, Matrix Heatmap using Seaborn. (le 29 septembre 2020) [https://vitalflux.com/correlation-heatmap-with-seaborn-pandas/]. (Consulté en 2021).

[25] Hesham Asem, (le 1 juillet 2019) « J-09- Feature Selection ». [https://www.youutube.com/watch?v=njXh6qYdpfs&list=PL6-3IRz2XF5X_9JeJh1xeciAbkijvc09k&index=9]. (Consulté en 2020).

[26] how-to-prepare-data-for-machine-learning (2013). [https://www.machinelearningmastery.ru/how-to-prepare-data-for-machine-learning/].(Consulté en 2020).

[27] Hesham Asem (le 26 juin 2019). "J-04-Data Cleaning" , [https://www.youtube.com/watch?v=awTU_lDQDYw]. (Consulté en 2020).

[28] Heart Disease Dataset. [Valable en line], [https://archive.ics.uci.edu/ml/datasets/heart+disease], (Consulté en 2020).

[29] Aanshi Gupta, Shubham Yadav, Shaik Shahid, Venkanna U. "HeartCare : IoT based heart disease prediction system", 2019 International Conference on Information Technology (ICIT).

[30] Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain and Preeti Nagrath, "Heart disease prediction using machine learning algorithms", IOP Conference Series : Materials Science and Engineering 2021.

[31] Archana Singh, Rakesh Kumar, "Heart Disease Prediction Using Machine Learning Algorithms", 2020 International Conference on Electrical and Electronics Engineering (ICE3-2020).

[32] Md. Nahiduzzaman, Md. Julker Nayeem, Md. Toukir Ahmed, "Prediction of Heart

Disease Using Multi-Layer Perceptron Neural Network and Support Vector Machine", 4th International Conference on Electrical Information and Communication Technology (EICT), December 2019.

[33] Yohan C. (le 19 novembre 2020). « Qu'est-ce que l'algorithme ? ». [https://datasciencetest.com/knn], (Consulté en 2021).

[34] Algorithme des k voisins les plus proches (KNN). [https:\\www.isnbreizh.fr/nsi/activity/algoRefKnn/index.html]. (Consulté en 2021).

[35] LE CŒUR ET LA CIRCULATION SANGUINE. (le 29 mars 2011), [http://www.lycee-sainte-cecile.com/sites/resources/files/Biologie-Physiopathologie/diaporama\%20LE\%20CUR\%20ET\%20LA\%20CIRCULATION\%20SANGUINE.pdf]. (Consulté en 2021).

[36] « Prévention des maladies cardiovasculaires », (Genève, 2007). [https://www.who.int/publications/list/cardio-pocket-guide-fr.pdf?ua=1], (Consulté en 2021).